An AI-Guided Framework Reveals Conserved Rules Governing microRNA Strand Selection

Author list

Dalton Meadows[1,2],

Hailee Hargis[1,2],

Amanda Ellis[1,2],

Heewook Lee[1],

Marco Mangone[1,*]

Author affiliations

[1)] The Biodesign Institute at Arizona State University, 1001 S McAllister Ave, Tempe, AZ

[2)] School of Life Sciences, Arizona State University, 427 E Tyler Mall, Tempe, AZ

Author list footnotes

[*)] Corresponding author

Corresponding author and e-mail address

Marco Mangone; mangone@asu.edu

Abstract

MicroRNAs (miRNAs) are central regulators of gene expression, yet the fundamental question of how cells choose between the two strands (5p or 3p) of a miRNA duplex during biogenesis remains unresolved. Here, we present a comprehensive, experimentally grounded framework that decodes the logic of miRNA strand selection. Using *Caenorhabditis elegans* as a model system, we developed a high-throughput qPCR platform enabling precise quantification of strand usage across development and in specific somatic tissues. Our analysis reveals that strand selection is not stochastic but follows conserved, strand-specific rules integrated with cellular and developmental context. To uncover the molecular grammar underlying this decision, we used artificial intelligence to build a predictive machine learning model trained on experimentally validated data. This AI-driven model, integrating 77 biologically informed features, accurately predicts strand preference not only in nematodes but also across vertebrates, including humans, revealing compositional and structural biases that are conserved yet functionally repurposed. Our work establishes a new paradigm for understanding miRNA strand selection, combining large-scale experimentation with AI to reveal an ancient, conserved, and programmable layer of gene regulation.

Introduction

miRNAs are ~22-nucleotide non-coding RNAs that guide the RNA-induced silencing complex (RISC) to target mRNAs, primarily through base pairing within the 3′ untranslated regions (3′UTRs) [1]. Once bound, the miRNA–RISC complex (miRISC) typically represses translation of the target mRNA. Both miRNAs and their 3′UTR targets are often conserved across metazoans and play key roles in essential biological processes such as development and morphogenesis.

miRNAs have undergone several pronounced expansions in metazoan genomes [2], coinciding with evolving organismal complexity yet rarely corresponding with increases in protein-coding genes [2]. Eukaryotic genomes contain 100-2,000 distinct mature miRNAs in intergenic regions, introns, and in clusters. Following transcription in long primary miRNA transcripts (pri-miRNAs) by the RNA polymerase II, they are processed in the nucleus by the microprocessor complex into shorter hairpin structures known as precursor miRNAs (pre-miRNAs) [1]. The two strands composing the pre-miRNAs are named the 5p and the 3p strand, respectively. Pre-mRNAs are subsequently exported to the cytoplasm by Exportin-5, where they undergo further cleavage by the enzyme Dicer, producing mature miRNA duplexes [1].

The Hsc70-Hsp90 chaperone complex facilitates the ATP-dependent loading of the miRNA duplex into Argonaute, the protein that execute the silencing step. This complex is thought to induce conformational changes in Argonaute that enable duplex loading [3-5]. Once loaded, one strand of this duplex, the guide strand, is kept within the RISC, while the other, the passenger strand, is typically degraded. Both the 5p and the 3p strands

can be selected for miRISC loading and act as guide strands. It is currently unclear how this process, termed 'miRNA strand selection,' is executed in detail, which proteins are involved in this decision, and how the guide strand is ultimately chosen between the two strands. The predominant factors influencing miRNA strand selection are believed to be the specific 5' nucleotide sequences of each strand and the comparative thermodynamic stability exhibited by the two ends of the miRNA duplex [6]. Studies performed using the human Ago2 protein, an Argonaute protein member of the RISC complex, show that the pocket responsible for binding the 5' nucleotide of the guide strand exhibits a strong preference for Uracil nucleotides [7]. This aligns with the prevalent occurrence of Uracil as the 5' nucleotide in miRNA guide strands [2], whereas the passenger strand commonly initiates with a Cytosine at its 5' end.

Furthermore, the miRISC complex exhibits a preference for loading the end of the miRNA duplex with lower thermodynamic stability, although not all the time [8]. The initial four nucleotides on each end of the duplex are believed to establish this thermodynamic asymmetry, and in some cases, a slight difference in the number of hydrogen bonds can induce the preferential loading of the less stable strand *in vitro* [5].

Consistently, the 5' ends of guide strands are enriched with purines [2]. In comparison, the 5' ends of passenger strands are predominantly pyrimidine-rich in both human and fly miRNAs, contributing to the thermodynamic asymmetry observed between the two ends of the duplex. In *C. elegans*, single amino acid substitutions within the MID (G553R) and PIWI (G716E, S895F, and S939F) domains of the Argonaute protein ALG-1 invert the strand selection process for specific miRNAs, suggesting a potential role for this protein in the miRNA strand selection process [9]. This indicates that while these

mutations play a pivotal role, they do not consistently reverse strand selection for all miRNAs, suggesting that there may be additional internal or external factors involved in this process. Tissue specific accessory factors and their dosage could also be responsible of miRNA strand selection, altering strand choices based on tissue or developmental contexts, similarly to the role of the SR proteins in RNA splicing [10]. While thermodynamic and sequence features help explain strand choice, a comprehensive mechanistic framework has been lacking.

C. elegans is an ideal model for studying miRNA strand selection due to its compact 3′UTRome [11-14], small number of miRNAs (miRbase), fully mapped cell lineage [15], and well-characterized transcriptome. Here we developed a high-throughput qPCR platform to systematically profile miRNA strand selection events across *C. elegans* development and in somatic tissues. These data revealed conserved sequence and structural patterns, which we used to train a machine learning model that predicts strand selection with 83% accuracy and generalizes to human miRNAs.

Materials and methods

All DI water was obtained from either a Purelab Option system (ELGA LabWater) or an Avidity Science system (Duo) and autoclaved before use. Chemicals were ordered from Sigma Aldrich/Merck, VWR, or Thermo Fisher Scientific in molecular biology grade and stored according to company instructions. Gateway™ LR Clonase II Plus™ enzyme mixes, SuperScript™ III Reverse Transcriptase, and PowerTrack™ SYBR Green Master Mix were purchased from Invitrogen. Primers for the HiTmiSS assay were ordered from

Life Technologies. Poly(U) Polymerase and restriction enzymes used in validation were ordered from New England Biolabs. The pCFJ150 plasmid was a gift from Erik Jorgensen. The OP50-1 *E. coli*, N2 *C. elegans* and EG6699 *C. elegans* strains were purchased through the Caenorhabditis Genetics Center.

Maintenance of *C. elegans* strains

All *C. elegans* strains were cultured on NGM plates seeded with OP50-1 bacteria. All transgenic lines produced for experiments described in this study were maintained at 24°C. When bacteria on each NGM plate had been cleared, *C. elegans* strains were transferred onto new plates via washing into a 15 ml Falcon tube using 15 ml M9 media (22 mM KH_2PO_4, 42 mM Na_2HPO_4, 85.5 mM NaCl, 1 mM $MgSO_4$, sterilized by autoclave) and pelleted for 3 minutes at 400*g*. The M9 media was then aspirated to a volume of 500 ml and the animals were resuspended via gentle pipetting before being moved onto fresh, seeded NGM plates.

Synchronization of *C. elegans* strains

The EG6699 strain [16] was maintained at 24°C. To synchronize animals for injection, confluent plates were bleached to isolate the embryos. Each plate was washed into a 15 ml Falcon tube with 15 ml DI water and pelleted for 3 minutes at 400*g*. The water was then aspirated, and the animals were incubated in 5 ml of bleaching solution (0.5M NaOH, 10% bleach) for 20 minutes with gentle end-over-end rotation. After bleaching,

the embryos were washed twice with DI water and moved onto fresh, seeded NGM plates. For use in the HiTmiSS assay, both N2 and *pges-1::GFP::alg-1::unc54 3´UTR C. elegans* strains were synchronized as described above.

Primers design for the HiTmiSS Assay

To quantify microRNAs in high throughput, we designed 380 unique forward primers for use in the qPCR amplification step of the HiTmiSS assay. These primers were designed to be complimentary to either the 5p strand or the 3p strand of 190 most expressed microRNAs in the *C. elegans* genome, corresponding to 75% of all *C. elegans* miRNA deposited in miRbase (release 22.1). We did not consider microRNAs whose existence is not supported by broad experimental evidence nor any microRNAs predicted to exist. Few miRNAs were ignored because their corresponding primers had a very low Tm, making it impossible to process using our pipeline. All primers were ordered in four 96-well microplates from Life Technologies pre-diluted at a concentration of 50 mM (**Supplemental Table S1**). We designed a universal reverse primer complementary to the stem-loop sequence of the stem-loop poly(A) primer used to prepare the cDNA for this qPCR reaction. 24 replicates of this primer were ordered from Life Technologies in a separate array, also at a concentration of 50 mM. Both forward and reverse primers were re-arrayed and diluted to a working concentration of 10 mM immediately prior to use.

***Singletons* and *Switchers* identification**

For each strand, we subtracted its relative quantity at one life stage versus the relative quantity of the complementary strand (i.e. for *let-7* 5p, we compared the abundance at embryo stage versus the abundance of *let-7* 3p at the embryo stage). This was repeated for each life stage at which the expression of that microRNA strand fell above our significance threshold (1 standard deviation above the mean of all quantified miRNAs at a given stage), and any microRNAs for which their expression was greater at each such life stage were deemed *Singletons*. We classified as *5p Singletons* microRNAs where the 5p strand was more expressed at each life stage than its corresponding 3p strand. We classified as *3p Singletons* microRNAs where the 3p strand was more expressed at each life stage than its corresponding 5p strand. were We classified as *Switchers* microRNAs which rose above our threshold of significance during at least one life stage, but whose expression levels did not display a consistent pattern of usage of one strand over the other.

HiTmiSS Assay

To collect sufficient total RNA for analysis of wild type miRNA, confluent NGM plates of N2 strains were synchronized and either washed and pelleted immediately (embryo stage) or incubated at 24°C for 15 hours (larval 1 stage), 25 hours (larval 2 stage), 33 hours (larval 3 stage), 42 hours (larval 4 stage), or 51 hours (young adult stage) before washing and pelleting. All washes were performed in 15 ml Falcon tubes using 15 ml DI water, and pellets were generated by centrifugation for 3 minutes at 400g, resulting in a

final pellet volume of 500 µL to 1 mL. Sufficient worms to generate this size pellet differed by life stage, requiring 12 (larval 4 and young adult stages), 18 (larval 2 and larval 3 stages), or 24 (embryo and larval 1 stages) starting confluent plates.

This pellet was then resuspended to a volume of 1.5 ml in ice cold Triton X-100 lysis buffer (150 mM NaCl, 1.0% Triton X-100, 50 mM Tris-HCl pH 8.0, 0.5 mM EDTA) and transferred to a 2 ml Spectrum tube pre-filled with 0.5 mm zirconium beads. Using a Benchmark BeadBug D1030 microtube homogenizer, each tube was homogenized for 90 seconds, incubated for 2 minutes on ice, then homogenized for an additional 60 seconds. The resultant lysate mixture was centrifuged for 3 minutes using a VWR benchtop mini centrifuge and the liquid fraction was transferred into a 1.5 ml Eppendorf microcentrifuge tube. To minimize non-aqueous contaminants, this lysate was centrifuged again in the same manner and transferred to a 15 ml Falcon tube pre-chilled on ice. To isolate total RNA, we used a DirectZol RNA MiniPrep Plus kit from Zymo Research, following kit instructions and performing the entire procedure at 4°C. Resultant RNA was eluted in 30 µL NF H_2O. To collect tissue-specific miRNA for analysis of the intestine, we performed protein immunoprecipitation followed by miRNA isolation. The fluorescent *pges1::gfp::alg-1::unc54 3′UTR C. elegans* strain previously prepared by our lab [17] was grown to confluence, synchronized, and pelleted as described above for each life stage, concurrent with N2 *C. elegans* as a negative control.

Each pellet was resuspended to 1.5 µl using our Triton X-100 lysis buffer with added protease inhibitor (two Roche cOmplete ULTRA mini EDTA-free protease inhibitor tablets per 50 ml lysis buffer) and transferred to a zirconium bead tube to be

homogenized as previously described. During this homogenization step, anti-GFP magnetic antibody beads (Chromotek GFP-Trap Magnetic Agarose) were prepared by diluting 25 µl bead slurry in 500 µl dilution buffer (150 mM NaCl, 50mM Tris-HCl pH 8.0, 0.5 mM EDTA) and rotating end-over-end at 4°C for 3 minutes.

The beads were pelleted using a benchtop centrifuge, and the buffer was aspirated using a magnetic tube rack. 500 µl dilution buffer was added again, and this dilution process was repeated twice. We then recovered aqueous lysate from the homogenized mixture as previously described and diluted it using 300 µl dilution buffer. This diluted lysate was transferred to a tube containing the prepared anti-GFP beads and left to rotate end-over-end at 4°C overnight. To clear unbound lysate, the beads were pelleted using a benchtop centrifuge and buffer was aspirated using a magnetic tube rack. 500 µl of wash buffer (150 mM NaCl, 50 mM Tris-HCl pH 8.0, 0.1% Triton X-100, 0.5 mM EDTA) was added, the beads were pelleted, and the buffer was aspirated using a magnetic tube rack. This wash step was repeated five additional times. The beads were then resuspended in 100 µl wash buffer and RNA was isolated using a DirectZol RNA MiniPrep Plus kit from Zymo Research, following kit instructions and performing the entire procedure at 4°C. Resultant RNA was eluted in 30 µL NF H_2O. The protocol for the HiTmiSS assay was adapted from Mei *et al.* [18]. We prepared our cDNA from each life stage for qPCR analysis via polyuridinylation using Poly(U) Polymerase (NEB) as per manufacturer instructions followed by a 20-minute heat inactivation at 65°C. We performed a first-strand reaction on this polyuridinylated RNA using Superscript III Reverse Transcriptase and stem-loop poly(A) primers under recommended conditions (Invitrogen). We then performed qPCR in high throughput using this cDNA, SYBR

Green master mix (Invitrogen), and our primers as described above. We used Applied Biosystems Quantstudio 3 and Quantstudio 7 qPCR machines to perform the screen. For each life stage, four 96-well plates were prepared with 40 μL reactions in each well, along with a positive control reaction containing 10 pg pDONR 221 ROG instead of cDNA and the necessary forward and reverse primers to amplify *gfp*. Each plate was then re-arrayed in triplicate into a 384-well plate using a Biomek FX^P Laboratory Automation Workstation for a total of three 10 μL reactions per microRNA strand per life stage. An additional plate was prepared with no cDNA or plasmid spike-in DNA as a non-template control. These qPCR reactions were carried out under standard SYBR Green reagent conditions on an Applied Biosystems QuantStudio 7. Relative miRNA levels were calculated using the $2^{-\Delta\Delta Ct}$ method (see below), and normalized using the spike-in positive control. These results were then plotted logarithmically as a heatmap. Whole worm data were ordered by type (*5p Singleton*, *3p Singleton*, or *Switcher*), while intestine data were clustered using average linkage. Both data were visualized using heatmapper.ca.

Cycle Threshold Analysis of the HiTmiSS Assay

The qPCR results containing the cycle threshold (Ct) values for each tested miRNA was exported in Microsoft Excel 2019 format. A comprehensive list of these data can be found in **Supplemental Table S2**. Ct values were averaged across experimental replicates, then subjected to a $2^{-\Delta\Delta Ct}$ analysis to calculate initial amounts of each microRNA in the sample, consistent with experimental standard [19]. The Ct value for any

given microRNA strand was subtracted from the Ct value of the non-template control (a qPCR reaction lacking any cDNA) for that microRNA strand. This was termed the ΔCt for that microRNA. The same was done for the positive control for that sample. Then, for each sample, this difference was subtracted from the difference obtained for that positive control, normalizing samples to the same positive control and giving us the ΔΔCt for that microRNA strand.

The inverse log base 2 of that ΔΔCt was then calculated to approximate the initial amount of that miRNA strand present in the cDNA relative to the positive control. This was deemed the $2^{-\Delta\Delta Ct}$ value for that miRNA strand. To determine whether a miRNA was considered present in a sample, we decided on a strict threshold of one standard deviation above the mean of all $2^{-\Delta\Delta Ct}$ values for that stage of development. MiRNAs quantified at levels above this threshold were considered significant for the purposes of this study. This same process was repeated to determine significance for our intestine tissue-specific HiTmiSS assay.

qPCR of microRNAs targeting overexpressed intestinal 3′UTRs

To quantify miRNA strand the levels when the target 3′ UTRs of those strands are overexpressed in the intestine, we used two *C. elegans* strains previously published by our lab [17]. Each expresses the promoter for intestine-specific gene *ges-1* fused to a dual color ROG construct (a gene encoding *mCherry* and *gfp* separated by a trans-spliceable operon) fused to the 3′ UTR of either *fox-1* or *hrp-2* [17]. Each of these strains was grown to confluence in 12 NGM plates in mixed stage and subjected to

quantification using the same approach as the HiTmiSS assay described above, though rather than quantifying every microRNA in these animals, we quantified the 5p and 3p strand of *miR-54*, *miR-60*, *miR-232*, and *miR-239b*. These quantifications were carried out in 96-well plates using cDNA (both prepared from our strains as well as control cDNA gathered from N2 *C. elegans*), SYBR Green master mix, and the appropriate forward and universal reverse primers for the miRNAs of interest, re arrayed from our standard HiTmiSS plates. Each 96-well plate contained two positive controls in *let-7* and *lin-4*, as well as a non-template negative control for each set of primers. These qPCRs were carried out in experimental triplicate under standard SYBR Green reagent conditions on an Applied Biosystems QuantStudio 3.

Relative miRNA levels were calculated using the $2^{-\Delta\Delta Ct}$ method as described above and normalized using the average of *let-7* and *lin-4* expression for each sample. These results were then averaged across two replicates and analyzed using Microsoft Excel 2019. Data were visualized as $\log_2$ of the change in expression from wildtype N2 negative control to either of our experimental strains. The threshold of significance was set at 1 as per standard analytical procedure.

Structural Analysis of microRNAs

The experimentally validated *5p* and *3p Singletons* were obtained using our HiTmiSS assay. Each entry included the entire hairpin sequence and its predicted secondary structure in dot-bracket notation [20] (**Supplemental Table S2)**. To focus specifically on mature miRNA regions, only the sequences of each mature miRNA

duplex were used for all downstream analyses. Since the *5p* and *3p Singleton* datasets contained different numbers of miRNAs, we balanced them before analysis by undersampling the larger group to match the sample size of the smaller class, ensuring unbiased statistical comparison between 5p and 3p strands.

We extracted 77 numerical features from each *5p* and *3p Singleton* (**Supplemental Table S3)**. The extracted features included: Base composition (fraction of A, U, G, C), motif presence (e.g. UGU, GAG, CCU), positional characteristics (first/last nucleotide identity), dinucleotide frequencies (e.g., di_GG, di_CC), GC content and GC-related asymmetry metrics (GC_diff, GC_ratio_5p3p), strand lengths, and number of mismatches (unpaired positions in the dot-bracket structure). The full list of the features and their descriptions are shown in **Supplemental Table S3**.

Feature calculations were implemented in Python using regular expressions and sequence statistics functions. Structural mismatch counts were determined by counting unpaired characters within the uppercase-aligned mature duplex. Comparative statistical analysis between 5p and 3p strands was conducted using Welch's two-sample t-test. All features were tested independently, and results were ranked by p-value. Features with $p < 0.05$ were considered statistically significant. Bar plots were generated for each significant feature using the mean ± standard deviation across strands.

miRNA strand selection ML model

To predict dominant strand selection (5p vs. 3p) in miRNA hairpins, we utilized ChatGPT GPT-4o (OpenAI) to assist in developing and training a Random Forest classifier based on 77 biologically curated features, organized into functional categories capturing distinct structural and sequence-related aspects of miRNA hairpins. These features capture thermodynamic stability, base-pairing patterns, and positional relationships within the miRNA hairpin. The model was trained on our validated *5p* and *3p Singletons* identified using our HiTmiSS assay, and feature values were computed using custom Python scripts that parse dot-bracket RNA structures and extract both absolute and relative measures for each strand. The weights of the features were optimized using known *Singletons*. The final model incorporates the following features, divided into five categories. The first category is the *Dinucleotide Composition* features, which quantify the frequency of specific two-nucleotide motifs within the 5p and 3p regions, capturing local sequence patterns that can influence strand selection and structural flexibility. The second category is the *First/Last Nucleotide Bias* features, which record the identity of the first and last nucleotides in each strand, reflecting known preferences in Argonaute loading where terminal nucleotides strongly impact strand choice. The third category is the *GC Content and Entropy* features, which assess the global nucleotide composition and thermodynamic diversity within each strand, including overall GC percentages, GC ratios between strands, and Shannon entropy measures that reflect structural variability and rigidity. The fourth category is the *Base-pairing Properties* features, which describes the fraction of nucleotides engaged in base pairing within the 5p and 3p regions, providing a measure of local hairpin stability and asymmetry. Finally, the *General Structural* features, which include the overall hairpin

length and the length difference between the 5p and 3p strands, capture broader geometrical properties that may indirectly affect miRNA maturation and strand retention (**Supplemental Figure S1**). The complete set of features, their description, and the weight used are shown in **Supplemental Table S3**. The model showed high performance in distinguishing 5p and 3p usage from a pool of known *Singletons* as a test dataset (*5p Singleton*: accuracy 100%, precision 75%, recall 100%; *3p Singleton:* 82% accuracy, 100% precision, 82% recall) (**Supplemental Figure S2**). The ML model uses a Random Forest classifier implemented in scikit-learn (v1.3.1). Model hyperparameters were optimized via grid search with 5-fold stratified cross-validation, tuning: n_estimators: [100, 250, 500], max_depth: [5, 10, None], min_samples_split: [2, 5, 10], class_weight: [‘balanced’, None]. The final selected parameters were: RandomForestClassifier (n_estimators=250, max_depth=None, min_samples_split=2, class_weight='balanced'). To prevent overfitting and enhance interpretability**,** recursive feature elimination (RFE) with cross-validation was used to rank features by importance. Features with near-zero variance and multicollinearity (|Pearson's r| > 0.9) were removed. The model consistently retained ~45–55 features, depending on the species subset. Performance was assessed on independent held-out test sets not used during training. Metrics included: Accuracy: 0.85 (*C. elegans*), 0.83 (human), Precision (5p/3p)**:** 0.87 / 0.80, Recall (5p/3p)**:** 0.89 / 0.78, F1 Score (5p/3p)**:** 0.88 / 0.79**,** AUROC**:** 0.91 (*C. elegans*), 0.89 (human). Class balance was maintained using SMOTE oversampling to correct for the slight enrichment of 5p strands in the source datasets. We confirmed that the model generalizes well by testing it on separate datasets (not used during training or tuning), including across species (e.g., training in *C. elegans*, testing in humans).

ML model predictions in human miRNAs

Human miRNA expression data were obtained from a harmonized dataset from the human miRNA tissue atlas [21] containing strand-specific Reads Per Million Mapped (RPMM) values across healthy human tissues. To restrict the analysis to non-diseased samples, we only considered entries where the Biotype column was labeled as ”tissue” “EV”, or “cell_line” (**Supplemental Table S6**), representing the average expression values across biological replicates of healthy tissues (total=337 datasets). For each miRNA duplex, we extracted the expression levels corresponding to the 5p and 3p strands, based on strand-specific annotations. We assigned strand dominance by summing RPMM values across all healthy tissue samples for the 5p and 3p strands separately. The strand with the higher total expression was designated as the dominant strand. We compared these dominant strands assignments with the predicted dominant strand assigned by our ML model. Each prediction was considered correct if the model-assigned strand matched the strand exhibiting greater expression across healthy tissues. Evaluation was performed on the topmost highly expressed 10, 50, and 100 miRNAs ranked by the sum of 5p and 3p strands RPMM values. We also performed the analysis using the complete dataset (all).

Results

A novel HT qPCR approach to study miRNA biogenesis

Identifying miRNAs in somatic tissues is challenging due to their small size, low abundance, and short half-life. To overcome this, we developed a high-throughput qPCR method that reliably quantifies miRNA expression and distinguishes 5p from 3p strand usage.

This method is a high throughput adaptation and general improvement of an existing methodology [18], which uses the polyuridine polymerase enzyme to elongate a polyU tail to all RNA transcripts in a given sample. This step extends the length of each miRNA present in the sample, which is then used as a template in reverse transcriptase reactions using a specially engineered universal stem-loop reverse primer (SL-polyA) (**Main Fig. 1A**).

We named this method High-Throughput miRNA Strand Selection (HiTmiSS) assay. HiTmiSS assay enables rapid, precise quantification of the full miRNAome and accurately distinguishes 5p from 3p strand expression (**Main Fig. 1**).

We employed this assay to systematically quantify miRNA occurrence and the usage of 5p and 3p strands in known miRNAs throughout *C. elegans* development. We extracted total RNA from *C. elegans* at six developmental stages (embryos, larval stages 1–4, and young adults) and analyzed the miRNA transcriptome and strand selection for 190 miRNAs, representing 75% of all miRNAs annotated in miRbase, using 380 unique forward primers across four plates (**Supplemental Table S1**).

We performed 7,980 qPCR experiments independently, querying these 190 miRNAs throughout development in triplicate (1,140 experiments at each stage) (**Main Fig. 1B, Supplemental Fig. S3**). Using stringent filters (see Materials and Methods), we

detected the presence of 177 of these miRNAs (93.2%) in at least one developmental stage (**Supplemental Table S2**). The greatest number of miRNAs occur during L4 stage (81%), followed by young adult (75.5%), L3 (65%), L2 (56.8%), and lastly, Embryo and L1 (41.6% each) stages. This aligns with previous literature suggesting that many *C. elegans* miRNAs are tissue-specific, only becoming more abundant upon the development of those tissues later in the developmental cycle [22].

As previously reported, early expressed miRNAs such as the key regulators of embryonic lethality *miR-35* and *miR-51* families, were most abundant during embryonic and early larval stages [23]. In contrast, *let-7* and *lin-4*, which control adult phenotypes, peaked at the L4 and young adult stages [24]. Notably, 13 miRNAs, including *miR-791*, *miR-792*, and *miR-8206*, showed no significant expression, likely due to low abundance or lack of experimental validation.

Many miRNAs show consistent strand preference throughout development

Many miRNAs identified consistently use either the 5p or 3p strand as the guide strand, which we termed '*Singletons*', while others switch between 5p and 3p usage across different developmental stages. We named the second category '*Switchers*'.

The screen produced hits for 177 miRNAs, which we divided into *5p Singletons*, *3p Singletons*, or *Switchers* (**Main Fig. 1B and Supplemental Table S2**).

The largest group identified was miRNA *Switchers* (46.3%, 88 hits). These include members of the *miR-2* family and other regulators of developmental timing and stress responses [24,25]. The remaining hits were *Singletons*, with a notable bias: *3p Singletons*

(58 hits, 30.5%) outnumbered *5p Singletons* by nearly twofold. These include *miR-35* and *miR-51* (embryonic lethality), *miR-44* (stress/starvation), and *miR-58* (locomotion and dauer entry) [23,24,26]. Several abundant *3p Singletons*, including *miR-60*, *miR-70*, and *miR-90*, remain functionally uncharacterized but are linked to stress response [25,27,28].

5p Singletons were the least common group, with 31 hits (16.3%), including the well-studied families *let-7* and *lin-4*, which are key regulators of developmental timing (**Main Fig. 1B**). Also detected were *miR-51* (embryonic lethality) and *miR-64* (heat shock response) [23,28]. While some families (*let-7* and *miR-64*), consistently use the same strand, others span multiple categories. For instance, the *miR-51* family splits evenly between *5p* (*miR-51* to *-53*) and *3p Singletons* (*miR-54* to *-56*, *lsy-6*), showing no *Switchers*, highlighting that strand usage can diverge even within structurally similar family members.

Importantly, several hits produced by the HiTmiSS assay align with past findings on the strand preference of well-characterized miRNAs such as *let-7, lin-4* (known *5p Singletons*), and the *miR-35* family (known *3p Singletons*) [24]

To account for the strong strand preference observed among *Singletons*, we took the top 20 *5p Singletons* and the top 20 *3p Singletons* and compared the binding signature and nucleotide composition between the used strand versus the strand that gets discarded (**Main Fig. 1C, i and ii**). Interestingly, the pairing frequency does not vary between these miRNAs (**Main Fig. 1C,ii**), either in the 5p or 3p strand, except for the last few nucleotides at the 3'end (**Main Fig 1C, I, red boxes**) which are commonly unpaired in the mature duplex.

When comparing the used vs. the discarded strand (**Main Fig. 1C,** left panel vs right

panel), we detected no significant difference in the pairing signature at any position; however, the 5p strand, independently if is used or discarded, contains more unpaired nucleotides (**Main Fig. 1C,** red boxes). When narrowing our analysis to just the seed region (**Main Fig. 1C**, golden boxes), while there is higher pairing than any other position, there is no difference in binding frequency between the used and discarded strands, suggesting that other rules are also in place.

Next, we compared the nucleotide composition of these seed regions, hoping to identify signatures that could explain the discrepancy in miRNA strand selection in *Singletons*. We determined the likelihood of any given nucleotide from positions 1 through 10 in each miRNA strand in the top 20 *Singletons* in each category (5p and 3p) (**Main Fig. 1C, ii**). We detected a highly conserved uracil nucleotide at position 1, as previously shown by us and others [2], though this conservation dissolves from positions 2 through 10 (**Main Fig. 1C, ii**). Importantly, this uracil nucleotide is not equally conserved between *5p Singletons* and *3p Singletons*. 80% of *3p Singletons* start with this uracil nucleotide, while less than 40% occur in *5p Singletons*. This differential presence of this uracil may be caused by a non-stochastic pre-miRNA cleavage by the *C. elegans* DCR-1, which may possess a preference to cleave at this nucleotide.

The intestine miRNA population is small but highly interconnected

The vast majority hits detected by our HiTmiSS assay were *Switchers* (~46%). Many of them may indeed not be true *Switchers* but rather act as tissue- or stage-specific *Singletons*, potentially confounding the analysis. To address this, we refined our

approach by analyzing somatic tissues across development, focusing on the intestine, which is a large tissue with a well-characterized transcriptome [29,30].

We coupled our HiTmiSS assay with intestine-specific miRNA immunoprecipitation, as we have published in the past [17], focusing on the embryo vs young adult stages. More specifically, we forced the expression of the GFP-tagged *C. elegans* Argonaute ortholog *alg-1* in the intestine tissue using the intestine-specific promoter *ges-1p*, performed an α-GFP immunoprecipitation, recovered the miRNA strands loaded onto ALG-1, and used them as templates in our HiTmiSS assay (**Main Fig. 2A**).

We found that of the 96 miRNAs tested, 64 (68%) were present during at least one stage within the intestine tissue (**Main Fig. 2**). Of these, 58 (91%) were *Singletons*, while 6 (9%) were *Switchers* (**Supplemental Fig. S4A**).

This strongly differs from our whole worm data, reinforcing that miRNAs previously labeled as *Switchers* may truly be *Singletons* on a tissue-by-tissue basis, only appearing as *Switchers* once all tissues are analyzed together.

Interestingly, the proportion of *5p Singletons* to *3p Singletons* is not the same in the intestinal tissue as in the mixed tissue (**Supplemental Fig. S4A**). In our initial analysis, about two-thirds of *Singletons* were *3p Singletons*, yet in the intestine tissue, nearly two-thirds of *Singletons* are *5p Singletons* (**Supplemental Fig. S4A**).

Another surprise was the presence of intestine-specific *Switchers* (**Main Fig. 2A and Supplemental Fig. S4B**). This includes the *miR-51* family, well-characterized regulators of embryonic lethality [23]. In addition, we detected intestine tissue *Switchers* among the *miR-44, miR-58*, and *miR-71* families (**Supplemental Fig. S4B**). These miRNA families have diverse functions, from regulating developmental timing to starvation, hypoxia,

stress response, and dauer entry [25,28,31,32]. Importantly, these are intestine-specific phenotypes, suggesting that strand switching in these miRNAs may have functional significance by expanding 3'UTR targeting repertoires in the intestine.

stage. These strains expressed the promoter for a single miRNA fused to GFP, and live imaging was used to validate the expression of these miRNAs at the appropriate life stage in the intestine (**Main Fig. 2C**). All four strains displayed the expected fluorescence.

miRNA target presence doesn't account for strand usage

Our HiTmiSS array enables precise, strand-resolved quantification of miRNAs *in vivo.* We considered the possibility that miRNA targets might stabilize specific strands, potentially biasing detection in our HiTmiSS assay. In this model, only target-bound strands would be retained, while the unused strand would degrade, falsely suggesting strand preference. To test this, we overexpressed *fox-1* and *hrp-2* 3'UTRs, each containing multiple binding sites for typically discarded strands (e.g., *miR-60* 5p, *miR-232* 5p, *miR-239b* 5p), in intestinal tissue (**Supplemental Figure S5**). If target-induced stabilization were driving strand selection, we would expect increased detection of these 5p strands. However, strand usage remained unchanged compared to wild-type controls, with 3p strands (i.e. *miR-60* 3p, **Supplemental Figure S5C**) still predominating.

Taken together, these results indicate that our qPCR approach is not simply detecting stabilized miRNA strands due to the presence of targets. Rather, the relative

strand abundance measured by the HiTmiSS assay likely reflects genuine strand selection, rather than artifacts introduced by target availability.

***5p Singletons* and *3p Singletons* follow different rules for selecting strand usage**

Next, we studied the structural features associated with *Singletons*, which could explain the strict and consistent strand usage within these miRNAs. When comparing the used and the discarded strand, the pairing frequency varies significantly within each *Singleton*, revealing new conserved patterns (**Main Fig. 3**). Specifically, the strand used in *Singletons* often has an unpaired nucleotide at position 12, while the discarded strand does not (**Main Fig. 3A**, yellow boxes). This lack of pairing in position 12 in both *5p Singletons* and *3p Singletons* usage is intriguing. In humans, certain Argonaute proteins perform cleavage of target RNAs between positions 10 and 11, which corresponds to positions 10 and 11 along the bound miRNA strand by Watson-Crick pairing [33]. *C. elegans* ALG-1 is largely homologous to human Argonaute proteins, and it may retain this cleavage activity at similar target positions, leading to the degradation of the unused strand due to its Watson-Crick base pairing at this crucial position. In addition, we detected differences in the binding frequency of the first-position nucleotide in both *5p Singletons* and *3p Singletons* (**Main Fig. 3A,** blue boxes). *3p Singletons* also exhibit unique pairings at positions 18-19 (**Main Fig. 3**, white box) not present in the discarded strain. These patterns are most clearly seen in specific *Singleton* miRNAs, both *5p Singletons* such as *miR-52* and *3p Singletons* such as *miR-75* (**Main Fig. 3B**).

An updated machine learning (ML) model for miRNA strand selection

We wanted to extract patterns and features that make *Singletons* so unique and predispose them to always use one strand over the other. To do that, we used our wet bench *Singleton* data as a training set for a machine learning model designed to predict 5p versus 3p strand usage in miRNA hairpins. We hypothesized that by training such model we could deconvolve the unique intrinsic features that characterize *Singletons*. We opted for a Random Forest classifier, which we trained using our experimentally validated *Singletons* identified through our HITmiSS assay. This model incorporates 77 features, systematically organized into five functional categories (dinucleotide composition, first/last nucleotide bias, GC content and entropy, base pairing properties, and general structural features) that together capture critical aspects of miRNA mature duplex structure (**Supplemental Fig. S1–S2, Supplemental Table S3-S4, and Materials and Methods**). The model demonstrated strong performance in distinguishing 5p from 3p usage across a set of known miRNAs (**Supplemental Fig. S2, and Materials and Methods**).

***5p* and *3p Singletons* possess distinct structural features**

Our ML model built using our *Singleton* data identified several intrinsic features in these miRNAs. To generalize this model across all *C. elegans* miRNAs, and study how *Switchers* can instead bypass these rules and switch miRNA strand preference, we applied our ML model and conducted a comparative feature analysis across the entire

C. elegans miRNAome (374 miRNAs, miRbase).

We extracted and calculated all 77 features used in our ML model for these miRNAs (**Supplemental Table S4**), keeping our definition of *Singletons* very stringent to increase the signal (confidence score >80%).

As expected, we identified 54 *5p Singletons*, 52 *3p Singletons,* and 268 *Switchers* (**Main Fig. 4A-B** and **Supplemental Table S4**). Out of 77 extracted structural and sequence features, 13 exhibited statistically significant differences between the 5p and 3p strands ($p < 0.05$, Welch's t-test) (**Main Fig. 4B-E**). The most pronounced difference was observed in the identity of the first nucleotide of the 3p strand, which in *3p Singletons* was consistently a uracil ($p = 3.98 \times 10^{-}$□), followed by an enrichment of cytosine in the last position ($p = 5.69 \times 10^{-}$□) (**Main Fig. 4C-4E**). Additionally, 3p strands showed an enrichment for specific motifs, particularly GAG ($p = 2.51 \times 10^{-}$□) (**Main Fig. 4E**). We also observed compositional differences: 3p strands displayed increased levels of cytosine at the first position (not shown) and tended to be slightly shorter than their 5p counterparts ($p = 8.90 \times 10^{-3}$) (**Main Fig. 4D-E**).

Surprisingly, our ML model did not cluster all the *Switchers* around a confidence score of ~50%. Instead, we see a multi-modal distribution with several periodical clusters (**Main Fig. 4A)**. This suggests that across this category, structural and thermodynamic features alone cannot fully explain strand usage, and extrinsic factors could move the needle between 5p or 3p choice in these miRNAs, perhaps depending on tissue or developmental context.

We next wanted to study the intrinsic features identified by our ML model. For these experiments we used *Singletons* to test if altering these rules could convert them into

Switchers. In *let-7*, modifying nucleotide pairing at positions 11–12 (UP11-12) or changing the first base (5'U→C) reduced its predicted *5p Singleton* confidence by up to 17%, consistent with our model (**Main Figure 3C-D**). Similar results were observed in *miR-52*, *miR-75*, *lin-4*, and *miR-35*: introducing bulges near the center disrupted strand preference, while removing them reinforced it (**Supplemental Fig. S6**), highlighting the central region's role in strand selection.

The first nucleotide's identity and pairing also influenced strand prediction, though with variable impact across miRNAs (**Supplemental Fig. S6**). Mutating this position altered strand usage by less than 10% in *5p Single*tons but up to 30% in *3p Singletons*, mirroring HiTmiSS results and further supporting our strand selection model.

Strand usage bias differs by species

Our ML model identified several key intrinsic features that clearly separate *Singletons* from *Switchers* in *C. elegans*. Since many miRNA sequences are conserved across species, we hypothesized that the features identified by our ML model were selectively maintained throughout evolutionary time. To do that, we applied the same model to the complete human miRNAome (**Main Fig. 5**) (**Supplemental Table S5**). Strand usage in *C. elegans* was nearly balanced (51.3% 5p vs. 48.7% 3p), suggesting flexible regulatory potential. In contrast, human miRNAs showed a strong 5p bias (75.9% 5p), indicating a shift toward fixed strand preference (**Main Fig. 5A**).

We identified 13 significantly different features between species ($p < 0.05$, Welch's t-test), with striking enrichment of guanine content ($p = 5.62 \times 10^{-2}$) and GG

dinucleotides ($p = 1.94 \times 10^{-2}$) on human 5p strands (**Main Fig. 5B**). Humans also showed higher CCU motif presence and cytosine content in 3p regions, including di_CC ($p < 5 \times 10^{-2}$), suggesting evolved sequence or processing preferences. These trends are evident in conserved miRNAs such as *let-7* (**Main Fig. 4C**). We also extended predictions to other species (mouse, rat, frog, zebrafish, fly) to explore evolutionary patterns (**Supplemental Table S5**).

Our ML model can accurately predict human miRNA strand selection

Our *C. elegans* ML model was also capable to identify sequence elements and structural features in human miRNAs. To validate these results, we compared its predictions to expression data from the Human miRNA Tissue Atlas across 65 healthy tissues [21] **(Main Fig. 6)**. For the top 10 most expressed miRNAs, the model reached 90% accuracy, which is higher than in *C. elegans*. As more miRNAs were included, accuracy declined, reaching 59% across the full dataset, yet remained statistically significant ($p \approx 7.6 \times 10^{-}$) **(Main Fig. 6A)**.

To reduce noise from low-expression miRNAs, we focused on the 50 most abundant duplexes (**Main Fig. 6B**). The model correctly predicted the dominant strand in 33 cases, yielding 66% accuracy ($p \approx 0.0164$). Most errors involved 3p-dominant miRNAs misclassified as 5p, reflecting a model bias toward 5p predictions. These misclassifications often occurred when 5p and 3p strands had similar expression levels, suggesting either model uncertainty or that these miRNAs may function as *Switchers* in humans (**Supplemental Table S6**).

Notably, two misclassified miRNAs, *miR-182* and *miR-378a*, are known *Switchers in vivo* [34,35]. In other cases, mispredictions occurred when the model's confidence scores for 5p and 3p usage were nearly equal (±10%). For instance, *miR-199a* and *miR-199b*, involved in hypoxia and cardiac regulation [36], were predicted as 5p dominant despite showing 3p expression and a narrow confidence margin (~7%). Similarly, *miR-124*, critical for neurogenesis, was misclassified with a 6% confidence gap, despite clear 3p dominance [37].

Still the majority of miRNAs were accurately predicted across diverse human tissues, including key regulators with strong strand bias, indicating that our *C. elegans*-trained model generalizes well to vertebrates (**Main Fig. 6C and Supplemental Table S6**). For example, *miR-21*, which is a highly expressed oncomiR, was correctly predicted as 5p-dominant, with over 50-fold 5p vs. 3p expression [38]. Similarly, *let-7a* showed a 248-fold 5p preference and was correctly classified [39]. Other highly expressed miRNAs such *miR-1* and miR-*141* were also correctly predicted as 3p dominant.

Overall, these findings demonstrate that the model is capable of reliably predicting strand usage for miRNAs in healthy human tissues, particularly in cases where strand bias is strong. Errors were enriched among duplexes with more balanced expression levels or context-dependent strand usage, highlighting the complexity of predicting which strand will be used in cases of weak or dynamic regulation of strand selection.

Discussion

In this study, we investigated miRNA strand selection by integrating ~8,000 HiTmiSS qPCR assays in *C. elegans* with structural modeling and machine learning. This allowed classification into *5p Singletons*, *3p Singletons*, or *Switchers*, and informed a predictive model validated against human miRNA expression data.

Nearly half of miRNAs (46%) behave as *Switchers*, altering strand usage across development, indicating that strand selection is context-dependent rather than fixed (**Main Fig. 1)**. In contrast, key developmental regulators such as *let-7*, *lin-4*, and *miR-35* are *Singletons*, likely reflecting a need for consistent strand usage to ensure stable gene expression. Notably, *Singletons* tend to be more highly expressed (**Main Fig. 1B**), perhaps due to stronger structural or thermodynamic constraints. In the intestine, tissue-specific ALG-1 immunoprecipitation revealed a reversal of the developmental trend, with more *5p* than *3p Singletons* (**Main Fig. 2**). This suggests that strand selection is also shaped by tissue-specific factors, such as the cellular environment or cofactors. Since we examined only one somatic tissue, some miRNAs classified as *Switchers* in our experiments may in fact be tissue-specific *Singletons*.

A concern was that target mRNAs might stabilize one strand, creating an artificial strand preference in our HiTmiSS assay. To test this, we overexpressed 3'UTRs with binding sites for typically low-abundance strands in the intestine. Strand levels remained unchanged, indicating that HiTmiSS reflects true strand usage, not target-driven stabilization (**Supplemental Figure S5**).

When comparing the structural composition of each mature duplex, a series of patterns emerged. *5p Singletons* tend to have the first nucleotide unpaired to its corresponding strand in the duplex (**Main Fig. 3A**). This difference in pairing repeats in

the middle of the duplex and lowers toward the end (**Main Fig. 3A**). Contrarily, *3p Singletons* present a near-opposite trend, with the first nucleotide more often paired and the end portion of the sequence also paired, though the center of the sequence shows similar unpairing to *5p Singletons*. These results support earlier models linking 5' end thermodynamic instability to strand selection [42,43], but add positional detail by showing that asymmetry varies along the duplex. Both *5p* and *3p Singletons* show central unpairing, likely reflecting a conserved feature for Dicer or Argonaute engagement but not strand bias. Notably, *5p Singletons* show decreasing pairing from the seed to the 3' end, while *3p Singletons* display the opposite trend. These opposing patterns suggest that specific structural cues along the duplex, not just global stability, guide strand selection, pointing to evolved, position-dependent regulatory signals within miRNA architecture.

To expand this analysis, we extracted 77 structural and compositional features from our validated *Singletons* and trained a Random Forest classifier (ML model) to predict dominant strand usage. When comparing these features (**Main Fig 4**), we found several strand-specific differences, including a significantly higher occurrence of uracil at the first nucleotide in 3p strands and enrichment for GAG motifs.

We validated our predictions using *in silico* site-directed structural mutations in canonical miRNAs such as *let-7*, *lin-4*, and *miR-52*, *miR-75* (**Main Fig. 3C**, **Supplemental Fig. S4**). Disrupting features such as central bulges (**Main Fig. 3C, "UP11-12"**) or first nucleotide pairing (**Main Fig. 3C, "1U->C"**), altered predicted strand preference to some degree, further confirming their role in the selection process. Surprisingly, the degree to which they impact miRNA strand preference was

approximately equal, but when disrupted in tandem led to an even greater loss of predicted strand preference (**Main Fig. 3C, "Combined"**, **Supplemental Fig. S6**). In addition, these findings phenocopy the trends seen in our experimentally validated *Singletons* (**Main Fig. 3A**). Taken together, these results suggest that the structural and thermodynamic features identified in our wet bench validated *5p* and *3p Singletons* are indeed critical regulators of strand selection in nematodes.

Expanding this framework, we then applied our ML model to predict strand usage across the human miRNA dataset (**Main Fig. 5A**). We were particularly curious to test if the structural and thermodynamic features and motifs identified in *C. elegans* were conserved in vertebrates.

Compared to the nearly equal distribution of 5p and 3p usage in *C. elegans*, using the same cutoff of 80%, the vast majority of human miRNAs are identified as *Switchers* (831/854 predicted *Switchers*, 97.3%). However, within this class there is a clear bias toward 5p strands, with over 75% of predictions favoring 5p strand usage (**Main Fig. 5A**). This divergence likely reflects increased specialization in miRNA maturation and Argonaute loading in vertebrates, potentially driven by evolutionary selection for strand-specific targeting or stability.

The prevalent 5p bias in human miRNAs may reflect evolutionary pressures rather than just biochemical preference. As miRNA families expanded during vertebrate, particularly mammalian, evolution, selective forces likely favored stable strand usage to ensure consistent and non-redundant regulation [2]. This trend would promote *Singletons* over *Switchers*, reducing regulatory variability. The fixation of 5p strand usage may thus

represent stabilizing selection for predictable miRNA function, with structural features reinforcing 5p bias evolving alongside increasing organismal complexity.

In addition to this imbalance in *Singletons* vs *Switchers*, the nucleotide composition of their 5p and 3p strands is also different, with a significant enrichment of guanine content in human 5p strands and cytosine-rich motifs in 3p strands (**Main Fig. 5B**). These compositional biases are consistent with known differences in Argonaute protein specificity and may reflect lineage-specific constraints on strand functionality [44].

Importantly, these cross-species trends persisted in known miRNAs such as *let-7*, which was consistently predicted and validated as a *5p Singletons* both in *C. elegans* and in human, although at different confidence levels. This supports the robustness of our predictive ML model and highlights conserved features of miRNA processing, even in the presence of divergent strand preferences.

A surprising trend emerges when we compare our ML model prediction in human miRNAs to the human miRNA Tissue Atlas. Our ML model developed in *C. elegans* is capable of accurately predicting miRNA strand usage in humans (**Main Fig. 6**). This is unexpected, considering the vast evolutionary gap (~600 million years) between these two systems, which share less than 35% of nucleotide conservation in genes [45]. Despite this, the features we have identified seem to be evolutionarily conserved.

This suggests that miRNA strand selection is, at its core, a fundamental cellular process that has evolved very little over time, even as miRNA regulatory networks expanded in complexity with the increased cellular specialization. This becomes more evident when we overlay miRNA strand usage in human tissues of different origins. Here, the same miRNA always is predicted to use the same strand in other somatic

tissues (**Main Fig. 6C**). This invariance suggests that strand selection is most likely encoded by intrinsic properties of the hairpin itself, rather than by external, context-dependent cues. Such robustness implies that strand choice is not merely a byproduct of downstream regulatory needs, but a core mechanistic feature established early in miRNA biogenesis. We propose that during vertebrate evolution, particularly with the rise of tissue specialization, miRNA strand usage remained anchored to conserved structural determinants even as the broader miRNA regulatory network expanded and diversified. This mechanistic rigidity may stabilize gene regulatory outputs in the face of cellular complexity.

While our model shows that strand selection for many miRNAs can be accurately predicted based on intrinsic properties of the hairpin structure, the prevalence of *Switchers* strongly suggests the involvement of additional, context-dependent regulatory mechanisms. These may include tissue-specific cofactors, differential Argonaute availability, or post-transcriptional modifications influencing duplex stability or Argonaute loading preferences.

In conclusion, our findings illustrate the complexity of miRNA strand selection and reveal that this process is not governed solely by thermodynamic asymmetry. Instead, a combination of structural motifs, sequence biases, and cellular context dictate strand usage. Our work establishes a predictive framework for understanding strand selection but also opens new avenues for studying the evolution of miRNA regulatory logic across metazoans.

Data Availability

The raw data produced by the HiTmiSS assay are included in the **Supplemental Table S2**. The ML model is included as Supplemental Data and is also available upon request.

Funding

This work was funded by the National Institutes of Health 5R01GM118796 awarded to M.M.

Acknowledgments

We thank Dr. Murugan for assistance in HiTmiSS assay optimization.

Author Contributions

D.M. and M.M. designed the experiments. D.M. performed the experiments. D.M. and M.M. prepared the ML model and performed the bioinformatics analysis. H.L. supervised the development of the ML model. D.M., A.E. and H.H. performed the HiTmiSS assay experiments. D.M. and M.M. prepared the main and supplemental figures and supplemental tables. D.M. and M.M. wrote the manuscript. All authors have read and approved the manuscript.

Declaration of Interests

The authors have no conflicts of interest.

Figure titles and legends

Main Figure 1: HiTmiSS Assay and Analysis on Mixed Tissue. **(A)** HiTmiSS protocol for RNA extraction, extension, reverse transcription, and quantification in high throughput. Steps include (i) *C. elegans* growth and homogenization, (2) RNA extraction and isolation, (3) RNA polyuridinylation, extension, and reverse transcription (details shown in ii ; orange: mRNAs, green: polyU, purple: SL-polyA reverse universal primer), (4) addition of 5p or 3p miRNA-strand-specific forward primers, (5) loading into 384-well plates, and (6) quantification. **(B)** Heatmap of miRNA expression (n=3) in mixed tissue clustered by quantity, strand use, and developmental stage. Stages include embryo (E), larval 1 through 4 (L1-L4) and young adult (YA). The heatmap legend represents $\log_{10}$ of expression values for each miRNA strand normalized by spike-in GFP control at each stage from -4 (deep purple) to 4 (bright yellow). Black color indicates instances of data which fall below strict expression threshold (one standard deviation above mean of all expression values at that stage). **(C)** Heatmap of pairing frequency by position for the used (left) and discarded (right) strand of the 20 most highly quantified *5p* and *3p Singletons* in our mixed tissue dataset. (i) blue: paired; red: unpaired). (ii) Average pairing frequency, and logo plot of nucleotide frequency at each position displayed below (gold box mark the seed region).

Main Figure 2: Analysis of Intestine Tissue-Specific miRNAs. **(A)** Schematics of the intestine-specific GFP-tagged ALG-1 immunoprecipitation. **(B)** Heatmap of miRNA expression in intestine tissue clustered by strand and developmental stage. The heatmap legend represents $\log_{10}$ of expression values for each miRNA strand

normalized by spike-in GFP control at each stage from -4 (deep purple) to 4 (bright yellow). Black color indicates instances of data which fall below strict expression threshold (one standard deviation above mean of all expression values at that stage). E: embryo; YA: young adults. **(C)** Representative brightfield (left) and GFP (right) images of selected *C. elegans* strains expressing GFP under the control of promoters for miRNA known to be highly expressed in the intestine, either during embryo stage (top two) or young adult stage (bottom two). Arrows point to GFP expression in the intestine tissues in both embryos and young adult stages.

Main Figure 3: Structural and Thermodynamic Determinants of Strand Selection. **(A)** Pairing frequency by position of the used (green) and discarded (red) strand for the most highly quantified *5p Singletons* and *3p Singletons* in mixed tissue (n=20). Positions at which pairing frequency most differs are boxed, including the first nucleotide (light blue), center of the duplex (gold), and end of the strand (white). **(B)** Representative pri-miRNA hairpins for a *5p Singleton* (*miR-52*) and a *3p Singleton* (*miR-75*) demonstrating occurrence of trends in pairing as described above. Mature duplex regions highlighted in pink. **(C)** Changes in *Singleton* predictive confidence for the miRNA *let-7* as a result of *in silico* site-directed mutations. RNAfold secondary structures of pri-miRNA with mature 5p strand (blue) and mature 3p strand (red) from left to right: wildtype (WT), unpaired nucleotides at positions 11 and 12 via GG→UU substitution (UP11-12), first nucleotide U→C substitution (1U → C), combination of both prior mutations (Combined). Table summarizes the predicted strand preference by our ML model for each mutation. **(D)** Bar chart displaying probability as described above.

Main Figure 4: Analysis of Mature miRNA Duplex Features in *C. elegans*. (A) Frequency distribution of *5p Singleton* predicted confidence by our ML model across all miRNAs in *C. elegans*. *3p Singletons* (yellow) are defined as miRNAs falling below a predicted 5p confidence of 20%; *Switchers* are defined as miRNAs with a predicted 5p confidence between 20% and 80%; *5p Singletons* (red) are defined as miRNAs with a predicted 5p confidence of 80% or greater. **(B)** Predicted categorization of all *C. elegans* miRNAs by our ML model as per the above threshold. **(C)** Predicted categorization of *C. elegans* miRNAs for which uracil is the first nucleotide of the 3p strand. **(D)** Length difference (5p strand - 3p strand) for each category of *C. elegans* miRNAs. *5p Singletons* display significantly ($p<0.05$) longer 5p strands than 3p strands when compared either to *3p Singletons* or to *Switchers*. **(E)** Heatmap of 32 features including GC content, miRNA length, hydrogen bonding, nucleotide pairing, strand stability, and MFOLD-derived parameters enrichment in *5p* and *3p Singletons*. Heatmap values represent mean values of each feature across the miRNA categories. For binary features (i.e. 5p_has_AAU, which is either 0 or 1) the number represents the proportion (e.g. 0.3 = 30% of miRNAs in that category have an AAU motif in the 5p strand). For continuous features (i.e. gc_content, pairing_frac_5p), the number is the average value across all miRNAs in that category. For nucleotide composition features (i.e. 5p_A, 3p_U), the number reflects the fraction of that nucleotide within the strand. Features clustered by average linkage.

Main Figure 5: Comparative Analysis of Mature miRNA Duplex Features between *C. elegans* and *H. sapiens*. **(A)** Frequency distribution of predicted *5p Singleton* confidence across all miRNAs in *H. sapiens*. *3p Singletons* (yellow) are defined as miRNAs falling below a predicted 5p confidence of 20% or below; *Switchers* are defined as miRNAs with a predicted 5p confidence between 20% and 80%; *5p Singletons* (red) are defined as miRNAs with a predicted 5p confidence of 80% or greater. **(B)** Four most significant ($p<0.05$) structural differences between *H. sapiens* (left) and *C. elegans* (right) mature miRNA duplexes. From top left to bottom right: mean occurrence of guanine in the 5p strand (5p_G), mean occurrence of guanine-guanine dinucleotide in either mature strand (di_GG), mean occurrence of cytosine-cytosine dinucleotide in either mature strand (di_CC), binary classifier for presence of cytosine-cytosine-uracil trinucleotide in the 3p strand (3p_has_CCU). **(C)** Alignment of *C. elegans* (*cel*) *let-7* and five *H. sapiens* (*hsa*) *let-7* homologs shows differences in CCU occurrence within the 3p strand (red boxes). **(D)** Heatmap of differences in mean value for each of the 77 model features between each category of miRNA in *C. elegans* (c) and *H. sapiens* (h). For binary features (i.e. 3p_has_CCU, which is either 0 or 1) the number represents the proportion (e.g. 0.3 = 30% of miRNAs in that category have an CCU motif in the 3p strand). For continuous features (i.e. gc_content, au_content), the number is the average value across all miRNAs in that category. For nucleotide composition features (i.e. 5p_A, 5p_U), the number reflects the fraction of that nucleotide within the strand. Features clustered by average linkage.

Main Figure 6: Model Accuracy for the Human miRNA Tissue Atlas Dataset. **(A)** Model accuracy expressed as a percentage of correct predictions for four different subsets of *H. sapiens* miRNAs (total 854 mature duplexes). From left to right, ranked by total expression of mature 5p and 3p strand across all tissue types: predicted accuracy among top 10 (9 correct, 1 incorrect), top 50 (33 correct, 17 incorrect), top 100 (66 correct, 34 incorrect), and all (502 correct, 352 incorrect) miRNAs in this dataset. **(B)** Confusion matrix of the strand usage prediction model among the top 50 most highly expressed *H. sapiens* miRNAs. Summarizes differences in 5p and 3p strand usage between experimentally validated (miRNA Atlas) and model-derived (ML Model) data. **(C)** Heatmap of prediction accuracy, either correct (red) or incorrect (blue) for the top 50 most highly expressed *H. sapiens* miRNAs across 65 *H. sapiens* tissue types and organs. Data are clustered using average linkage.

Supplemental Information titles and legends

Supplemental Table S1: **Primers used in this study.**

Supplemental Table S2: **HiTmiSS assay results.**

Supplemental Table S3: **Features used in our ML model and their weights.**

Supplemental Table S4: **ML prediction in *C. elegans*.**

Supplemental Table S5: **ML prediction in vertebrates.**

Supplemental Table S6: **Comparison of ML model with miRNA tissue Atlas.**

References

1 Bartel, D. P. Metazoan MicroRNAs. *Cell* **173**, 20-51 (2018). https://doi.org:10.1016/j.cell.2018.03.006

2 Wolter, J. M. *et al*. Evolutionary patterns of metazoan microRNAs reveal targeting principles in the let-7 and miR-10 families. *Genome Res* **27**, 53-63 (2017). https://doi.org:10.1101/gr.209361.116

3 Pare, J. M. *et al*. Hsp90 regulates the function of argonaute 2 and its recruitment to stress granules and P-bodies. *Mol Biol Cell* **20**, 3273-3284 (2009). https://doi.org:10.1091/mbc.e09-01-0082

4 Nakanishi, K. Anatomy of RISC: how do small RNAs and chaperones activate Argonaute proteins? *Wiley Interdiscip Rev RNA* **7**, 637-660 (2016). https://doi.org:10.1002/wrna.1356

5 Meijer, H. A., Smith, E. M. & Bushell, M. Regulation of miRNA strand selection: follow the leader? *Biochem Soc Trans* **42**, 1135-1140 (2014). https://doi.org:10.1042/BST20140142

6 Medley, J. C., Panzade, G. & Zinovyeva, A. Y. microRNA strand selection: Unwinding the rules. *Wiley Interdiscip Rev RNA* **12**, e1627 (2021). https://doi.org:10.1002/wrna.1627

7 Frank, F., Sonenberg, N. & Nagar, B. Structural basis for 5'-nucleotide base-specific recognition of guide RNA by human AGO2. *Nature* **465**, 818-822 (2010). https://doi.org:10.1038/nature09039

8 Melendez, A. & Levined, B. in *WormBook* (ed The C. elegans Research Community) (WormBook, August 24, 2009).

9 Zinovyeva, A. Y., Veksler-Lublinsky, I., Vashisht, A. A., Wohlschlegel, J. A. & Ambros, V. R. Caenorhabditis elegans ALG-1 antimorphic mutations uncover functions for

Argonaute in microRNA guide strand selection and passenger strand disposal. *Proc Natl Acad Sci U S A* **112**, E5271-5280 (2015). https://doi.org:10.1073/pnas.1506576112

10 Caceres, J. F., Stamm, S., Helfman, D. M. & Krainer, A. R. Regulation of alternative splicing in vivo by overexpression of antagonistic splicing factors. *Science* **265**, 1706-1709 (1994). https://doi.org:10.1126/science.8085156

11 Mangone, M., Macmenamin, P., Zegar, C., Piano, F. & Gunsalus, K. C. UTRome.org: a platform for 3'UTR biology in C. elegans. *Nucleic Acids Res* **36**, D57-62 (2008). https://doi.org:10.1093/nar/gkm946

12 Mangone, M. *et al.* The landscape of C. elegans 3'UTRs. *Science* **329**, 432-435 (2010). https://doi.org:10.1126/science.1191244

13 Steber, H. S., Gallante, C., O'Brien, S., Chiu, P. L. & Mangone, M. The C. elegans 3' UTRome v2 resource for studying mRNA cleavage and polyadenylation, 3'-UTR biology, and miRNA targeting. *Genome Res* **29**, 2104-2116 (2019). https://doi.org:10.1101/gr.254839.119

14 Murari, E., Meadows, D., Cuda, N. & Mangone, M. A comprehensive analysis of 3'UTRs in Caenorhabditis elegans. *Nucleic Acids Res* (2024). https://doi.org:10.1093/nar/gkae543

15 Sulston, J. E. Neuronal cell lineages in the nematode Caenorhabditis elegans. *Cold Spring Harb Symp Quant Biol* **48 Pt 2**, 443-452 (1983). https://doi.org:10.1101/sqb.1983.048.01.049

16 Frokjaer-Jensen, C. *et al.* Single-copy insertion of transgenes in Caenorhabditis elegans. *Nat Genet* **40**, 1375-1383 (2008). https://doi.org:10.1038/ng.248

17 Kotagama, K., Schorr, A. L., Steber, H. S. & Mangone, M. ALG-1 Influences Accurate mRNA Splicing Patterns in the Caenorhabditis elegans Intestine and Body Muscle

Tissues by Modulating Splicing Factor Activities. *Genetics* **212**, 931-951 (2019). https://doi.org:10.1534/genetics.119.302223

18 Mei, Q. *et al.* A facile and specific assay for quantifying microRNA by an optimized RT-qPCR approach. *PloS one* **7**, e46890 (2012). https://doi.org:10.1371/journal.pone.0046890

19 Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) Method. *Methods* **25**, 402-408 (2001). https://doi.org:10.1006/meth.2001.1262

20 Lorenz, R. *et al.* ViennaRNA Package 2.0. *Algorithms Mol Biol* **6**, 26 (2011). https://doi.org:10.1186/1748-7188-6-26

21 Rishik, S., Hirsch, P., Grandke, F., Fehlmann, T. & Keller, A. miRNATissueAtlas 2025: an update to the uniformly processed and annotated human and mouse non-coding RNA tissue atlas. *Nucleic Acids Res* **53**, D129-D137 (2025). https://doi.org:10.1093/nar/gkae1036

22 Brosnan, C. A., Palmer, A. J. & Zuryn, S. Cell-type-specific profiling of loaded miRNAs from Caenorhabditis elegans reveals spatial and temporal flexibility in Argonaute loading. *Nat Commun* **12**, 2194 (2021). https://doi.org:10.1038/s41467-021-22503-7

23 Shaw, W. R., Armisen, J., Lehrbach, N. J. & Miska, E. A. The conserved miR-51 microRNA family is redundantly required for embryonic development and pharynx attachment in Caenorhabditis elegans. *Genetics* **185**, 897-905 (2010). https://doi.org:10.1534/genetics.110.117515

24 Alvarez-Saavedra, E. & Horvitz, H. R. Many families of C. elegans microRNAs are not essential for development or viability. *Curr Biol* **20**, 367-373 (2010). https://doi.org:10.1016/j.cub.2009.12.051

25 Garcia-Segura, L. *et al*. High-Throughput Profiling of Caenorhabditis elegans Starvation-Responsive microRNAs. *PloS one* **10**, e0142262 (2015). https://doi.org:10.1371/journal.pone.0142262

26 Garcia, S. M., Tabach, Y., Lourenço, G. F., Armakola, M. & Ruvkun, G. Identification of genes in toxicity pathways of trinucleotide-repeat RNA in C. elegans. *Nature structural & molecular biology* **21**, 712-720 (2014).

27 Kato, M., Kashem, M. A. & Cheng, C. An intestinal microRNA modulates the homeostatic adaptation to chronic oxidative stress in C. elegans. *Aging (Albany NY)* **8**, 1979-2005 (2016). https://doi.org:10.18632/aging.101029

28 Kagias, K., Podolska, A. & Pocock, R. Reliable reference miRNAs for quantitative gene expression analysis of stress responses in Caenorhabditis elegans. *BMC Genomics* **15**, 222 (2014). https://doi.org:10.1186/1471-2164-15-222

29 Blazie, S. M. *et al.* Alternative Polyadenylation Directs Tissue-Specific miRNA Targeting in Caenorhabditis elegans Somatic Tissues. *Genetics* **206**, 757-774 (2017). https://doi.org:10.1534/genetics.116.196774

30 Haenni, S. *et al.* Analysis of C. elegans intestinal gene expression and polyadenylation by fluorescence-activated nuclei sorting and 3'-end-seq. *Nucleic Acids Res* **40**, 6304-6318 (2012). https://doi.org:10.1093/nar/gks282

31 Pandit, A., Jain, V., Kumar, N. & Mukhopadhyay, A. PHA-4/FOXA-regulated microRNA feed forward loops during Caenorhabditis elegans dietary restriction. *Aging (Albany NY)* **6**, 835-855 (2014). https://doi.org:10.18632/aging.100697

32 Brunquell, J., Snyder, A., Cheng, F. & Westerheide, S. D. HSF-1 is a regulator of miRNA expression in Caenorhabditis elegans. *PloS one* **12**, e0183445 (2017). https://doi.org:10.1371/journal.pone.0183445

33 Nakanishi, K. When Argonaute takes out the ribonuclease sword. *J Biol Chem* **300**, 105499 (2024). https://doi.org:10.1016/j.jbc.2023.105499

34 Krishnan, K. *et al.* MicroRNA-182-5p targets a network of genes involved in DNA repair. *RNA* **19**, 230-242 (2013). https://doi.org:10.1261/rna.034926.112

35 Krist, B., Florczyk, U., Pietraszek-Gremplewicz, K., Jozkowicz, A. & Dulak, J. The Role of miR-378a in Metabolism, Angiogenesis, and Muscle Biology. *Int J Endocrinol* **2015**, 281756 (2015). https://doi.org:10.1155/2015/281756

36 Rane, S. *et al.* Downregulation of miR-199a derepresses hypoxia-inducible factor-1alpha and Sirtuin 1 and recapitulates hypoxia preconditioning in cardiac myocytes. *Circ Res* **104**, 879-886 (2009). https://doi.org:10.1161/CIRCRESAHA.108.193102

37 Makeyev, E. V., Zhang, J., Carrasco, M. A. & Maniatis, T. The MicroRNA miR-124 promotes neuronal differentiation by triggering brain-specific alternative pre-mRNA splicing. *Mol Cell* **27**, 435-448 (2007). https://doi.org:10.1016/j.molcel.2007.07.015

38 Rhim, J., Baek, W., Seo, Y. & Kim, J. H. From Molecular Mechanisms to Therapeutics: Understanding MicroRNA-21 in Cancer. *Cells* **11** (2022). https://doi.org:10.3390/cells11182791

39 Brueckner, B. *et al.* The human let-7a-3 locus contains an epigenetically regulated microRNA gene with oncogenic function. *Cancer Res* **67**, 1419-1423 (2007). https://doi.org:10.1158/0008-5472.CAN-06-4074

40 Kim, C., Ye, Z., Weyand, C. M. & Goronzy, J. J. miR-181a-regulated pathways in T-cell differentiation and aging. *Immun Ageing* **18**, 28 (2021). https://doi.org:10.1186/s12979-021-00240-1

41 Pakravan, K. *et al.* MicroRNA-100 shuttled by mesenchymal stem cell-derived exosomes suppresses in vitro angiogenesis through modulating the mTOR/HIF-1alpha/VEGF signaling axis in breast cancer cells. *Cell Oncol (Dordr)* **40**, 457-470 (2017). https://doi.org:10.1007/s13402-017-0335-7

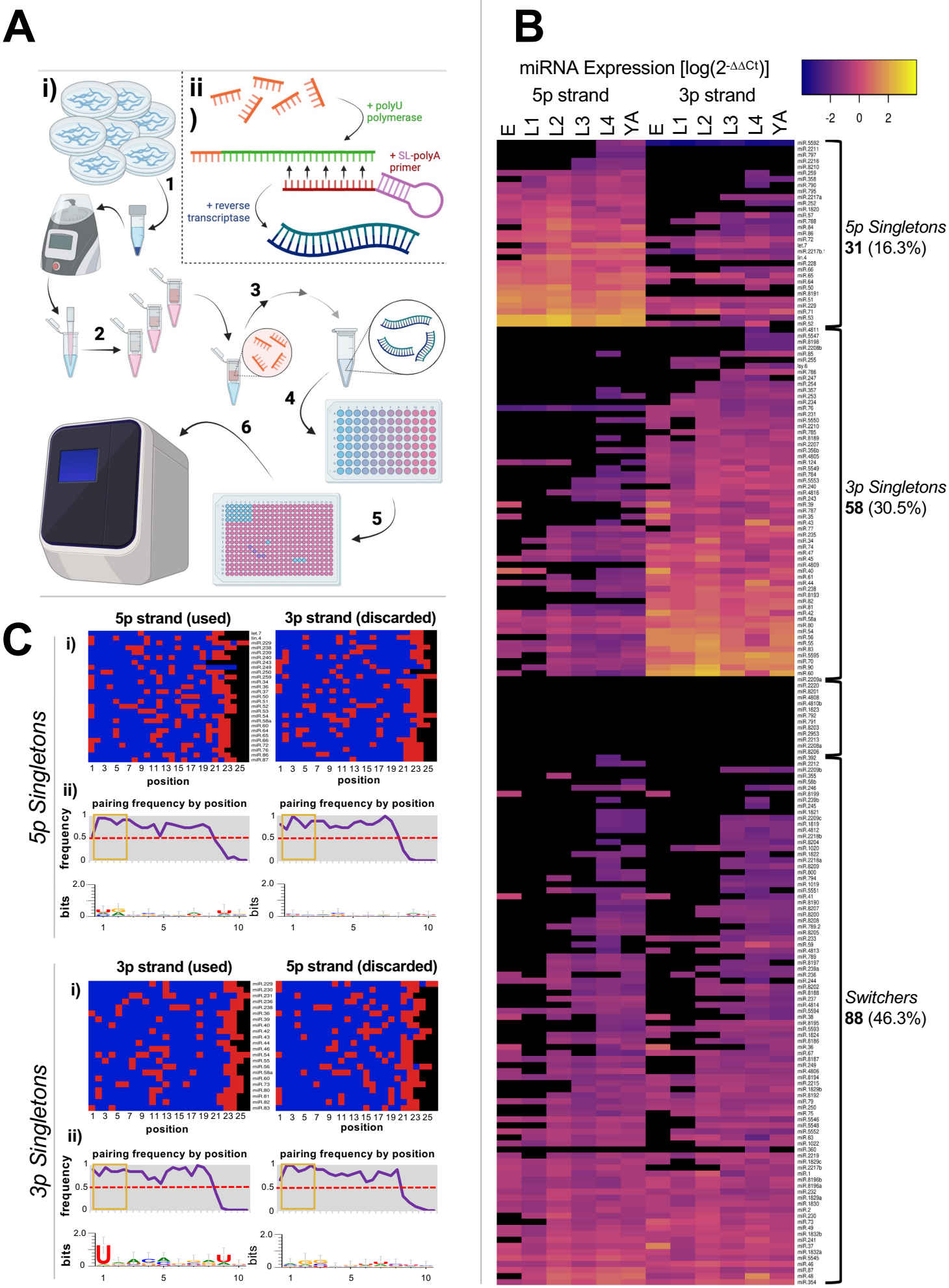
42 Khvorova, A., Reynolds, A. & Jayasena, S. D. Functional siRNAs and miRNAs exhibit strand bias. *Cell* **115**, 209-216 (2003). https://doi.org:10.1016/s0092-8674(03)00801-8

43 Schwarz, D. S. *et al.* Asymmetry in the assembly of the RNAi enzyme complex. *Cell* **115**, 199-208 (2003). https://doi.org:10.1016/s0092-8674(03)00759-1

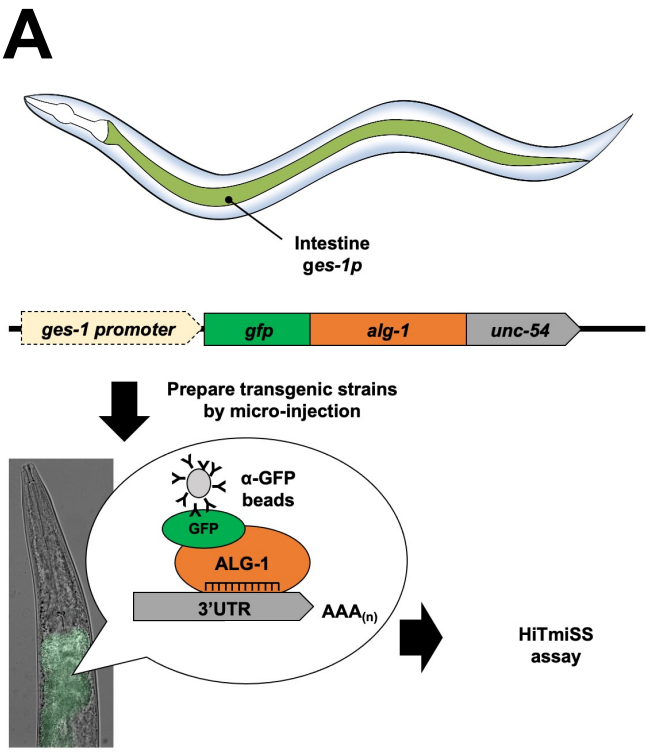
44 Kotagama, K. *et al.* Catalytic residues of microRNA Argonautes play a modest role in microRNA star strand destabilization in C. elegans. *Nucleic Acids Res* **52**, 4985-5001 (2024). https://doi.org:10.1093/nar/gkae170

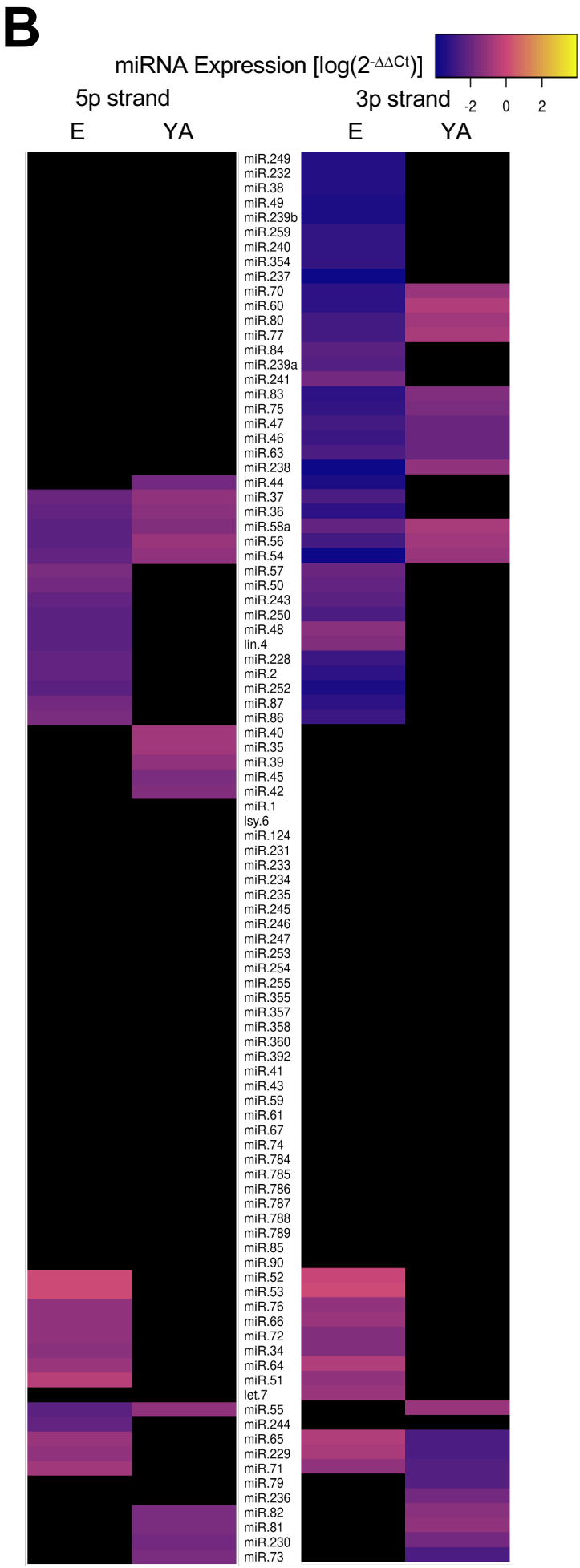
45 Stein, L. *et al.* WormBase: network access to the genome and biology of Caenorhabditis elegans (vol 29, pg 82, 2001). *Nucleic Acids Research* **29**, 1012 (2001).

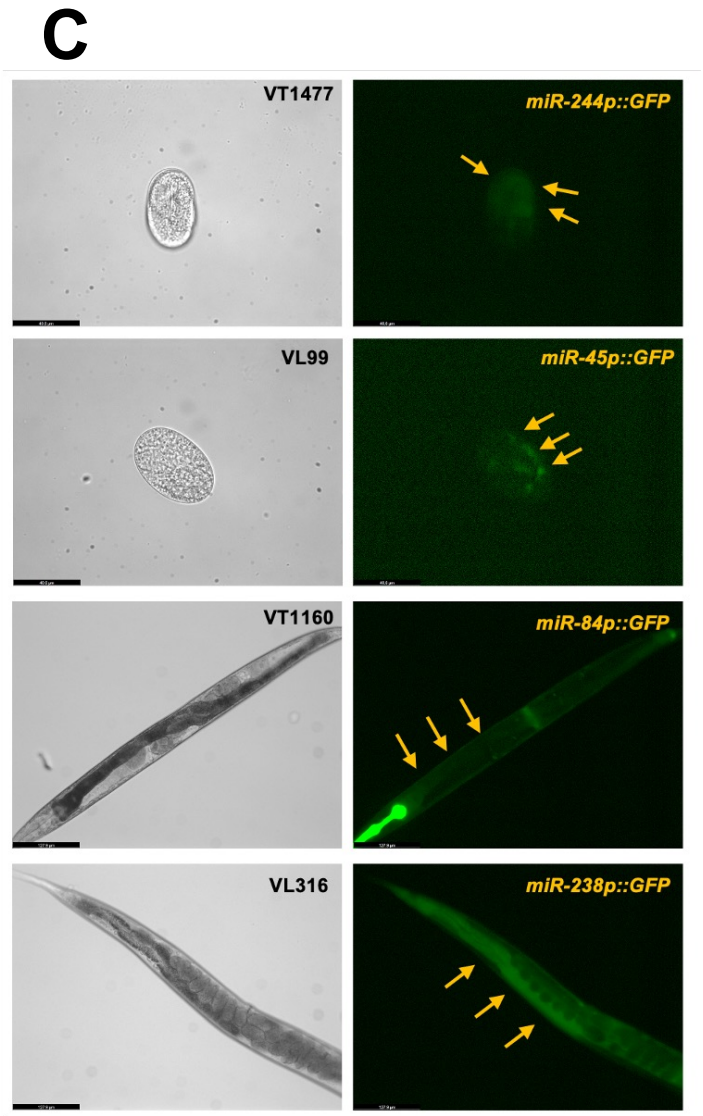
Main Figure 1



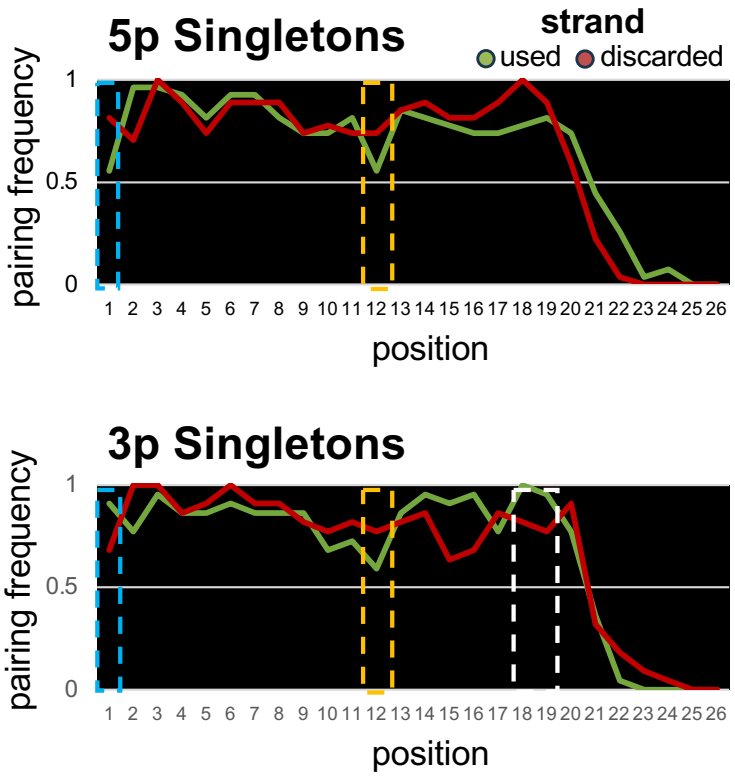
Main Figure 2



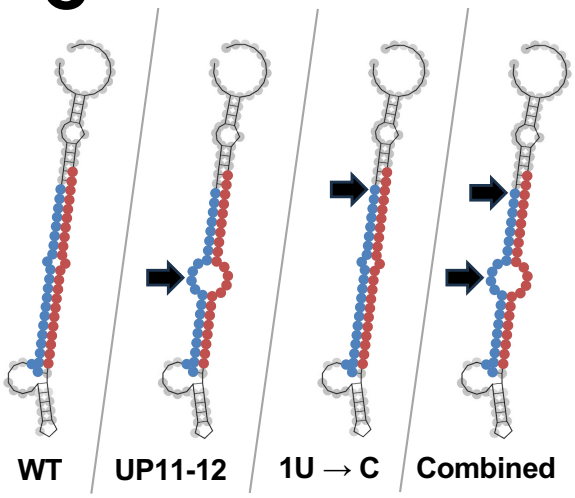




A

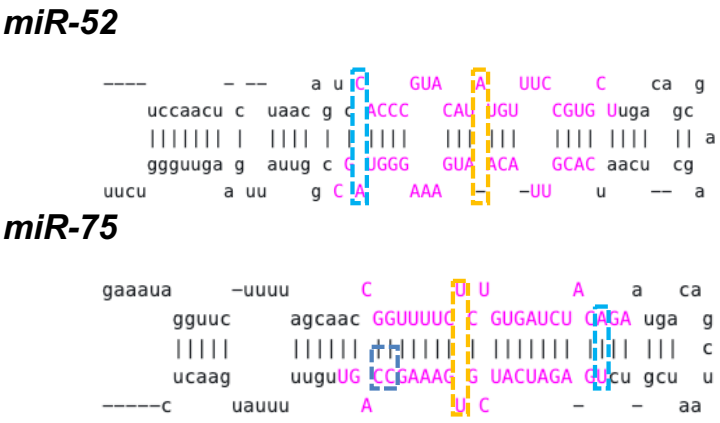


C

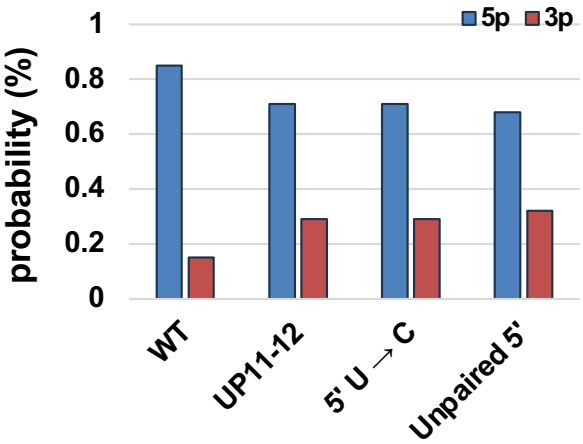


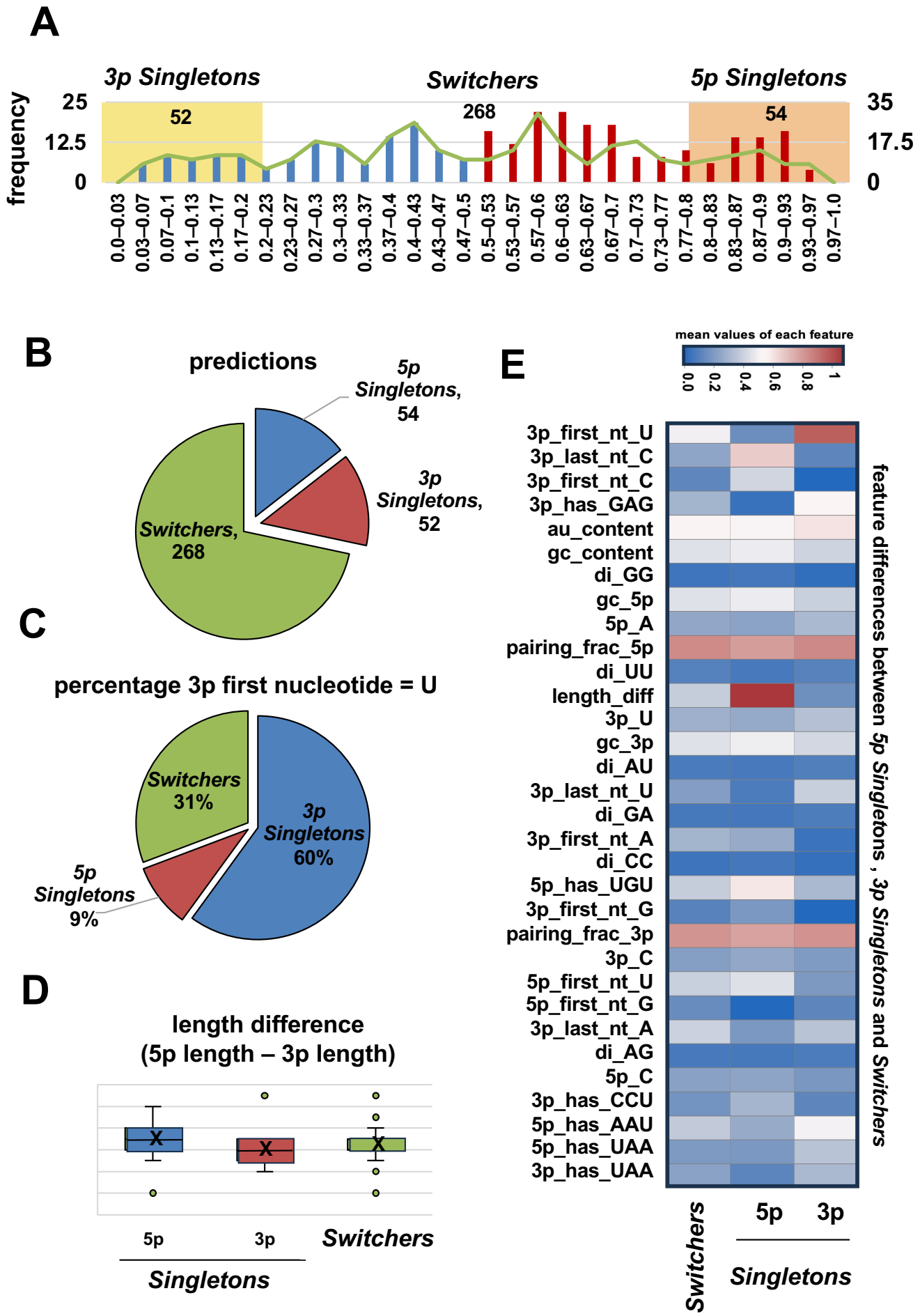
Variant	5p	3p
WT	0.85	0.15
UP11-12	0.71	0.29
5'U → C	0.71	0.29
Combined	0.68	0.32

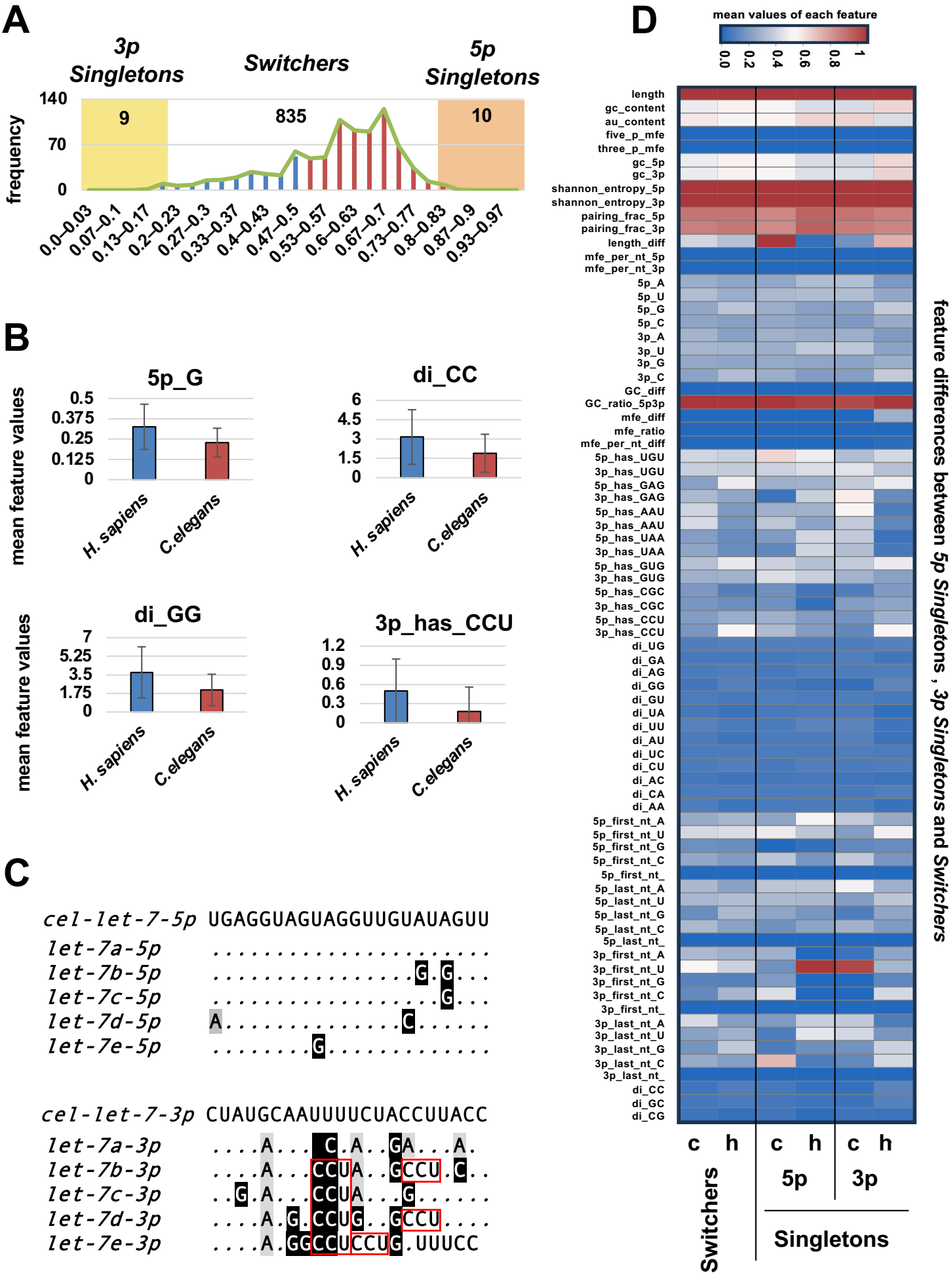
B

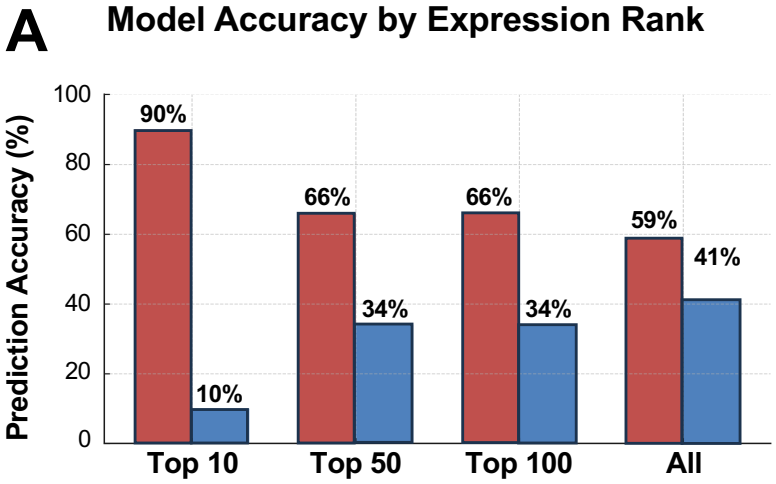


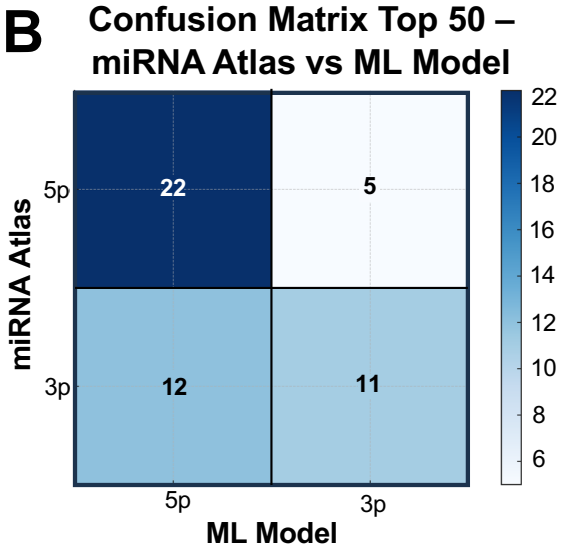
D

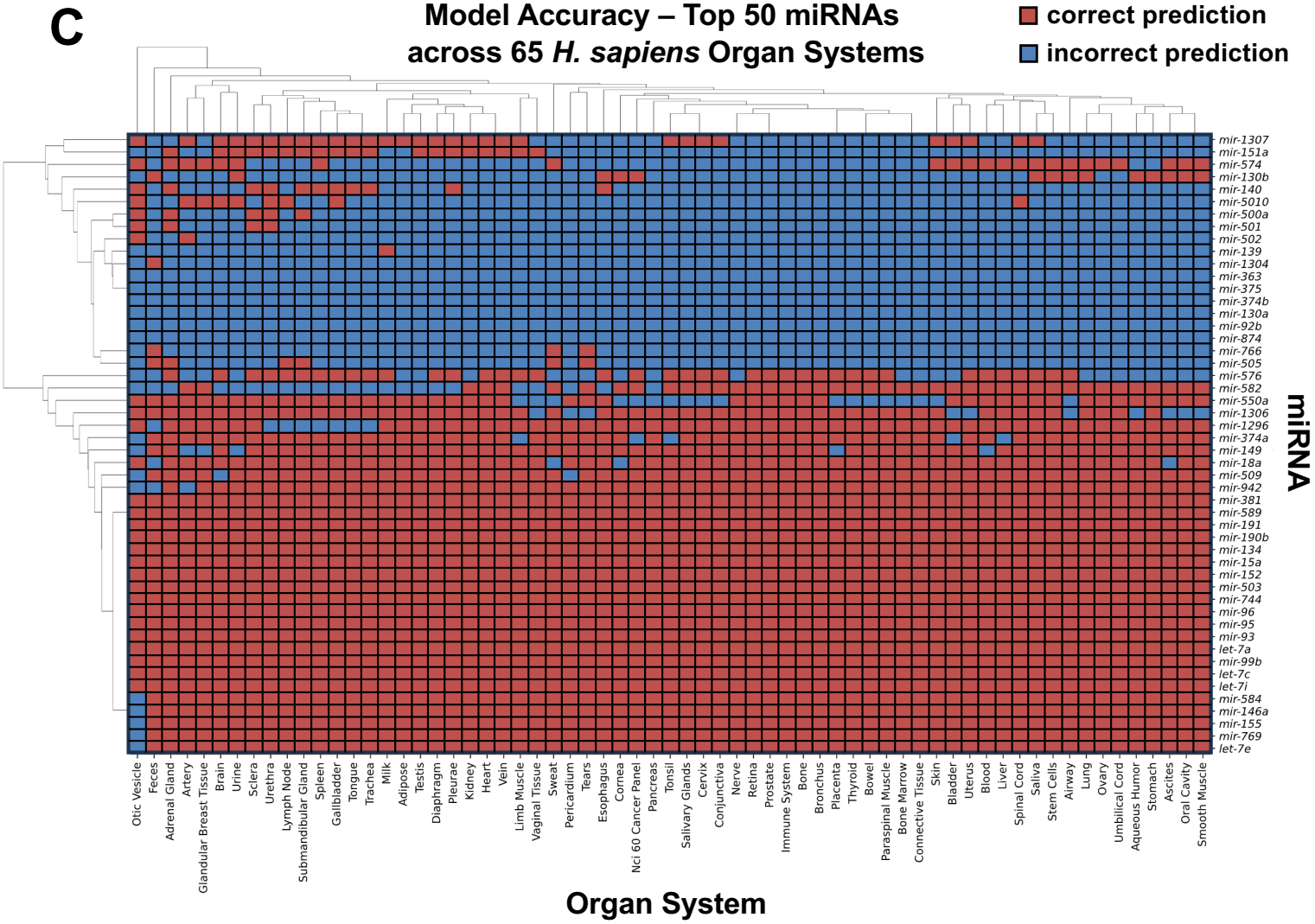



A
3p Singletons
Switchers
5p Singletons
frequency
25
12.5
0
35
17.5
0
52
268
54
0.0–0.03
0.03–0.07
0.07–0.1
0.1–0.13
0.13–0.17
0.17–0.2
0.2–0.23
0.23–0.27
0.27–0.3
0.3–0.33
0.33–0.37
0.37–0.4
0.4–0.43
0.43–0.47
0.47–0.5
0.5–0.53
0.53–0.57
0.57–0.6
0.6–0.63
0.63–0.67
0.67–0.7
0.7–0.73
0.73–0.77
0.77–0.8
0.8–0.83
0.83–0.87
0.87–0.9
0.9–0.93
0.93–0.97
0.97–1.0
B
predictions
5p Singletons, 54
3p Singletons, 52
Switchers, 268
C
percentage 3p first nucleotide = U
Switchers 31%
3p Singletons 60%
5p Singletons 9%
D
length difference
(5p length – 3p length)
5p
3p
Switchers
Singletons
E
mean values of each feature
0.0
0.2
0.4
0.6
0.8
1
3p_first_nt_U
3p_last_nt_C
3p_first_nt_C
3p_has_GAG
au_content
gc_content
di_GG
gc_5p
5p_A
pairing_frac_5p
di_UU
length_diff
3p_U
gc_3p
di_AU
3p_last_nt_U
di_GA
3p_first_nt_A
di_CC
5p_has_UGU
3p_first_nt_G
pairing_frac_3p
3p_C
5p_first_nt_U
5p_first_nt_G
3p_last_nt_A
di_AG
5p_C
3p_has_CCU
5p_has_AAU
5p_has_UAA
3p_has_UAA
feature differences between 5p Singletons , 3p Singletons and Switchers
Switchers
5p
3p
Singletons


A
3p Singletons
Switchers
5p Singletons
frequency
140
70
0
9
835
10
0.0–0.03
0.07–0.1
0.13–0.17
0.2–0.23
0.27–0.3
0.33–0.37
0.4–0.43
0.47–0.5
0.53–0.57
0.6–0.63
0.67–0.7
0.73–0.77
0.8–0.83
0.87–0.9
0.93–0.97
B
mean feature values
5p_G
0.5
0.375
0.25
0.125
0
di_CC
6
4.5
3
1.5
0
di_GG
7
5.25
3.5
1.75
0
3p_has_CCU
1.2
0.9
0.6
0.3
0
H. sapiens
C.elegans
C
cel-let-7-5p UGAGGUAGUAGGUUGUAUAGUU
let-7a-5p
let-7b-5p
let-7c-5p
let-7d-5p
let-7e-5p
cel-let-7-3p CUAUGCAAUUUUCUACCUUACC
let-7a-3p
let-7b-3p
let-7c-3p
let-7d-3p
let-7e-3p
D
mean values of each feature
0.0
0.2
0.4
0.6
0.8
1
length
gc_content
au_content
five_p_mfe
three_p_mfe
gc_5p
gc_3p
shannon_entropy_5p
shannon_entropy_3p
pairing_frac_5p
pairing_frac_3p
length_diff
mfe_per_nt_5p
mfe_per_nt_3p
5p_A
5p_U
5p_G
5p_C
3p_A
3p_U
3p_G
3p_C
GC_diff
GC_ratio_5p3p
mfe_diff
mfe_ratio
mfe_per_nt_diff
5p_has_UGU
3p_has_UGU
5p_has_GAG
3p_has_GAG
5p_has_AAU
3p_has_AAU
5p_has_UAA
3p_has_UAA
5p_has_GUG
3p_has_GUG
5p_has_CGC
3p_has_CGC
5p_has_CCU
3p_has_CCU
di_UG
di_GA
di_AG
di_GG
di_GU
di_UA
di_UU
di_AU
di_UC
di_CU
di_AC
di_CA
di_AA
5p_first_nt_A
5p_first_nt_U
5p_first_nt_G
5p_first_nt_C
5p_first_nt_
5p_last_nt_A
5p_last_nt_U
5p_last_nt_G
5p_last_nt_C
5p_last_nt_
3p_first_nt_A
3p_first_nt_U
3p_first_nt_G
3p_first_nt_C
3p_first_nt_
3p_last_nt_A
3p_last_nt_U
3p_last_nt_G
3p_last_nt_C
3p_last_nt_
di_CC
di_GC
di_CG
feature differences between 5p Singletons , 3p Singletons and Switchers
c h c h c h
5p
3p
Switchers
Singletons


A
Model Accuracy by Expression Rank
Prediction Accuracy (%)
100
80
60
40
20
0
90%
10%
66%
34%
66%
34%
59%
41%
Top 10
Top 50
Top 100
All


B
Confusion Matrix Top 50 – miRNA Atlas vs ML Model
miRNA Atlas
5p
3p
22
5
12
11
5p
3p
ML Model
22
20
18
16
14
12
10
8
6


C
Model Accuracy – Top 50 miRNAs across 65 H. sapiens Organ Systems
correct prediction
incorrect prediction
miRNA
mir-1307
mir-151a
mir-574
mir-130b
mir-140
mir-5010
mir-500a
mir-501
mir-502
mir-139
mir-1304
mir-363
mir-375
mir-374b
mir-130a
mir-92b
mir-874
mir-766
mir-505
mir-576
mir-582
mir-550a
mir-1306
mir-1296
mir-374a
mir-149
mir-18a
mir-509
mir-942
mir-381
mir-589
mir-191
mir-190b
mir-134
mir-15a
mir-152
mir-503
mir-744
mir-96
mir-95
mir-93
let-7a
mir-99b
let-7c
let-7i
mir-584
mir-146a
mir-155
mir-769
let-7e
Otic Vesicle
Feces
Adrenal Gland
Artery
Glandular Breast Tissue
Brain
Urine
Sclera
Urethra
Lymph Node
Submandibular Gland
Spleen
Gallbladder
Tongue
Trachea
Milk
Adipose
Testis
Diaphragm
Pleurae
Kidney
Heart
Vein
Limb Muscle
Vaginal Tissue
Sweat
Pericardium
Tears
Esophagus
Cornea
Nci 60 Cancer Panel
Pancreas
Tonsil
Salivary Glands
Cervix
Conjunctiva
Nerve
Retina
Prostate
Immune System
Bone
Bronchus
Placenta
Thyroid
Bowel
Paraspinal Muscle
Bone Marrow
Connective Tissue
Skin
Bladder
Uterus
Blood
Liver
Spinal Cord
Saliva
Stem Cells
Airway
Lung
Ovary
Umbilical Cord
Aqueous Humor
Stomach
Ascites
Oral Cavity
Smooth Muscle
Organ System